# Deep Neural Attention for Misinformation and Deception Detection

by

**Rahul Mishra**

A dissertation submitted in partial satisfaction of
the requirements for the degree
PHILOSOPHIAE DOCTOR (PhD)


Faculty of Science and Technology
Department of Electrical Engineering and Computer Science
June 2021

University of Stavanger
N-4036 Stavanger
NORWAY
www.uis.no



ISBN 978-82-8439-015-4
ISSN 1890-1387

PhD Thesis UiS no. 597

*This*
*Thesis Is*
*Affectionately Dedicated*
*To My Encouraging Parents, & Sisters*
*&*
*To My Loving Wife, & Child*

# Preface

This dissertation is submitted in partial fulfillment of the requirement of the degree of Philosophiae Doctor (PhD) at the University of Stavanger, Norway. The research has been conducted at the University of Stavanger, Norway from June 2018 to June 2021, including research visits to computer science department of ETH Zurich, Switzerland (January 2020 to April 2020) and Leibniz University Hannover, Germany (June 2019 to June 2019).

This dissertation consists of a collection of 5 research articles, which are included within the dissertation after required transformations and realignments to adhere to the requisite format. The content of the originally published articles has been kept intact.

*Rahul Mishra, June 2021*

# Abstract

At present the influence of social media on society is so much that without it life seems to have no meaning for many. This kind of over-reliance on social media gives an opportunity to the anarchic elements to take undue advantage. Online misinformation and deception are vivid examples of such phenomenon. The misinformation or fake news spreads faster and wider than the true news [32]. The need of the hour is to identify and curb the spread of misinformation and misleading content automatically at the earliest.

Several machine learning models have been proposed by the researchers to detect and prevent misinformation and deceptive content. However, these prior works suffer from some limitations: *First*, they either use feature engineering heavy methods or use intricate deep neural architectures, which are not so transparent in terms of their internal working and decision making. *Second*, they do not incorporate and learn the available auxiliary and latent cues and patterns, which can be very useful in forming the adequate context for the misinformation. *Third*, Most of the former methods perform poorly in early detection accuracy measures because of their reliance on features that are usually absent at the initial stage of news or social media posts on social networks.

In this dissertation, we propose suitable deep neural attention based solutions to overcome these limitations. For instance, we propose a claim verification model, which learns embddings for the latent aspects such as author and subject of the claim and domain of the external evidence document. This enables the model to learn important additional context other than the textual content. In addition, we also propose an algorithm to extract evidential snippets out of external evidence documents, which serves as explanation of the model's decisions. Next, we improve this model by using improved claim driven attention mechanism and also generate a topically diverse and non-redundant multi-document fact-checking summary for the claims, which helps to further interpret the model's decision making. Subsequently, we introduce a novel method to learn influence and affinity relationships among the social media users present on the propagation paths of the news items. By modeling the complex influence relationship among the users, in addition to textual content,

we learn the significant patterns pertaining to the diffusion of the news item on social network. The evaluation shows that the proposed model outperforms the other related methods in early detection performance with significant gains.

Next, we propose a synthetic headline generation based headline incongruence detection model. Which uses a word-to-word mutual attention based deep semantic matching between original and synthetic news headline to detect incongruence. Further, we investigate and define a new task of incongruence detection in presence of important cardinal values in headline. For this new task, we propose a part-of-speech pattern driven attention based method, which learns requisite context for cardinal values.

# Acknowledgements

First and foremost, I'd like to thank and express my gratitude to my PhD supervisor Prof. Dr. Reggie Davidrajuh, who is a constant source of encouragement and wisdom. I am very grateful for his patience, indispensable suggestions and emotional support. I would also like to thank my co-supervisor Prof. Dr. Krisztian Balog for his invaluable suggestions and constructive feed-backs.

I express my sincere gratitude to the head of the department at IDE, Dr. Tom Ryen for his kind support and guidance, throughout my PhD journey.

My heartfelt thanks to Prof. Dr. Thomas Hofmann for hosting me as a visiting researcher at Data Analytics Lab, ETH, Zürich.

I would also like to thank my collaborators and co-authors Prof. Dr. Markus Leippold (University of Zürich), Assoc. Prof. Dhruv Gupta (NTNU) and Shuo Zhang (Bloomberg U.K.) for their valuable contributions and fruitful discussions.

I thank all the faculty members and staff at IDE for creating an enabling research environment.

I would also like to thank Assoc. Prof. Dr. Antorweep Chakravorty and Prof. Dr. Trygve Christian Eftestøl for giving me an opportunity to take part in pedagogical activities in their courses besides routine research.

I would also like to thank all my colleagues and friends at IDE, especially Dr. Jayachander Surabirayala, Trond Linjordet and Dhanya Therese Jose. For their support, help and many technical and interesting discussions.

I am forever indebted to my caring parents, Ram Prakash and Aarti Mishra, and my sisters Anita, Sunita and Alka for their unwavering faith in me. Last but not least, I am forever grateful to my wife, Swati Mishra and Son, Anant Mishra, for their love, unequivocal support throughout and great patience at all times. Without them, none of this would have been indeed possible.

# List of Abbreviations

| | |
|---|---|
| NLP | Natural Language Processing |
| ML | Machine Learning |
| MTL | Multi-Task Learning |
| MSE | Mean Squared Error |
| SGD | Stochastic Gradient Descent |
| SSL | Semi-Supervised Learning |
| SVM | Support Vector Machine |
| BPTT | Back-Propagation Through Time |
| CNN | Convolutional Neural Network |
| GAN | Generative Adversarial Network |
| LSTM | Long Short-Term Memory |
| Bi-LSTM | Bidirectional Long Short-Term Memory |
| NN | Neural Network |
| ReLU | Rectified Linear Unit |
| Tanh | Hyperbolic Tangent |
| RNN | Recurrent Neural Network |
| LSTM | Long Short Term Memory |
| GRU | Gated Recurrent Unit |
| ROC | Receiver Operating Characteristic |
| AUC | Area Under the ROC Curve |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| LDA | Latent Dirichlet Allocation |
| BERT | Bidirectional Encoder Representations from Transformers |

# Contents

# List of Papers

The following papers are included in this thesis:

- **Paper I**

  **SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection**
  Rahul Mishra, Vinay Setty
  Published in 2019, The 9th ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.

- **Paper II**

  **Fake News Detection using Higher-order User to User Mutual-attention Progression in Propagation Paths**
  Rahul Mishra
  Published in 2020 The 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, Washington, USA

- **Paper III**

  **Generating Fact Checking Summaries for Web Claims**
  Rahul Mishra, Dhruv Gupta, Markus Leippold
  Published in proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text. Punta Cana,

Dominican Republic

- **Paper IV**

  **MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching**
  Rahul Mishra, Piyush Yadav, Remi Calizzano, Markus Leippold
  Published in the proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, Florida, USA.

- **Paper V**

  **POSHAN: Cardinal POS Pattern Guided Attention for News Headline Incongruence**
  Rahul Mishra, Shuo Zhang
  Accepted ( in publication ) in the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Queensland, Australia

# Chapter 1

# Introduction

*But thou hast only the right to work, but none to the fruit thereof. Let not then the fruit of thy action be thy motive; nor yet be thou enamored of inaction.*

*Lord Krishna*
*Srimad Bhagavad Gita*
*Chapter 2, Verse 47*

We are living in a digital world, which is seamlessly connected. The technological revolutions in the field of communication and information science, catalysed by the inception of social media platforms and hand held communication contrivances are unprecedented. These technological breakthroughs and advancements have led to a paradigm shift in the way people consume information. This expeditious magnification in digital information consumption has also resulted in some loopholes and shortcomings withal, such as proliferation of misinformation/fake news and deceptive contents. The recent surge in misinformation can be largely attributed to the rise of technology mediated communication mediums [12, 69, 40, 77].

**'Bogus' AP tweet about explosion at the White House wipes billions off US markets**

The FBI and SEC are to launch investigations after more than £90bn was temporarily wiped off the US stock market when hackers broke into the Twitter account of the Associated Press and announced that two bombs had exploded at the White House, injuring Barack Obama.


Figure 1.1: A fake news wiped out billions in the US stock market [a]

[a] https://www.telegraph.co.uk/finance/markets/10013768/Bogus-AP-tweet-about-explosion-at-the-White-House-wipes-billions-off-US-markets.html

## 1.1 Amiss Content

Fake news or misinformation is a kind of misrepresentation or dissemination of amiss information to be deceived and get economic or political benefits [1]. The spread of misinformation is not a modern phenomenon but the precise beginning is hard to contemplate and it might predate the historical recordings. Around 31 BC, after the death of Julius Caesar, his adopted son Octavian conspired against Mark Antony, the loyal general of Julius and spread many rumors and misinformation regarding his character and association with Cleopatra, which resulted in the defeat of Mark Antony [2].

[1] https://en.wikipedia.org/wiki/Fake_news

[2] https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6

In February, Tedros Adhanom Ghebreyesus, director-general of the World Health Organization, declared that Covid-19 was not the only public health emergency the world was facing — we were also suffering from an "infodemic" of fake medical news. "Fake news spreads faster and more easily than this virus," he said, "and is just as dangerous."

Figure 1.2: Excerpt from a news article, which covers the speech of director-general of WHO on seriousness of infodemic.

In the present era however, misinformation has become more catastrophic due to it's scale and reach [40, 21]. Some of the most striking examples of adversity caused by misinformation are the USA stock market crash[3], which costed $130 billion in stock value due to a fake tweet regarding explosion at the White House and the spread of a large amount of misinformation in the event of the Covid19 pandemic, which could lead to serious social and fatal health effects. The world health organization (WHO) has coined a new term "Infodemic[4]" for rapid spread of deceptive, fabricated and misleading content.

Deceptive contents are appetizing social media posts or news items that are by design exaggerated to appear more sensational and attractive to the users. Incongruent news headlines and clickbaits are some of the most popular form of deceptive contents. In a typical clickbait scenario, users are tempted by a very catchy news headline and once they click to read the rest of the news; They are dismayed to see that they have been deceived by the title and the news does not match its title. This is a serious problem pertaining to user experience and user friendliness.

In this thesis, we call both misinformation/fake news and deceptive contents together as **"Amiss Content"**, for sake of brevity.

[3] https://business.time.com/2013/04/24/how-does-one-fake-tweet-cause-a-stock-market-crash/

[4] https://www.who.int/news-room/feature-stories/detail/immunizing-the-public-against-misinformation

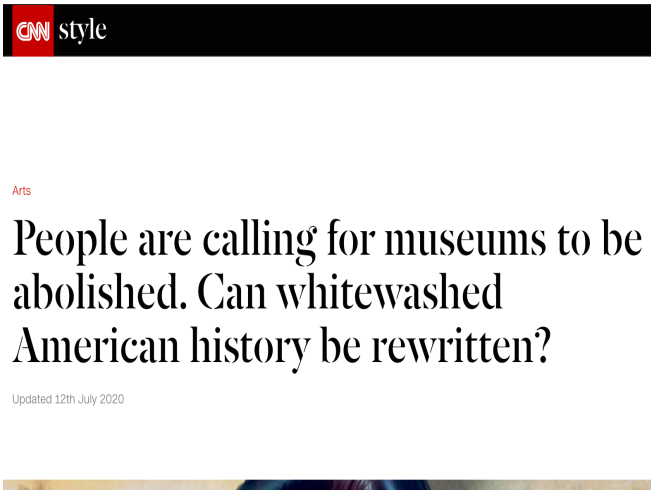



Figure 1.3: An excerpt from a CNN News article in which the headline has been sensationalized by portraying people's demand for the removal of a suspicious statue from a museum as if people are demanding the dismantling of all museums [a].

[a] https://edition.cnn.com/style/article/natural-history-museum-whitewashing-monuments-statues-trnd/index.html

## 1.2 Automated Amiss Content Detection

Keeping track of the veracity of humongous amount of social media posts and news items manually, seems beyond the bounds of possibility. Manual fact checking or veracity prediction is very tedious and time taking task[24], therefore automated misinformation detection or fact verification is the need of the hour. Machine learning methods have been used extensively in curbing the misinformation and deceptive contents online by major social media players and search engines such as Twitter, Facebook and Google. The research community has shown a lot of interest and dedication towards maintaining the sanity of the Web by proposing techniques and models to detect the amiss content online. These proposed and bench-marked methods are utilized by social media platforms and content driven industries such as digital news media. The other major objective of the automation of the amiss content detection is to aid and equip web users with tools, which can help them decide the truthfulness of a news piece or social media post themselves.

The main objective of this dissertation is to curb the spread of misinformation and deceptive content on the Web by automated detection and generation of the explanation in form of convincing evidences. To this end, by and large our research focuses on ***(a)*** to learn and utilize the contextual representations for the latent aspects or side information pertaining to the social media posts or news items such as users, who interacted with the news on social platform, the source or publisher of the news, and propagation path of the news on social network etc., ***(b)*** to extract an evidential excerpt or summary for explaining the verdict of the model, and ***(c)*** to come up with novel and suitable neural attention mechanisms to cater to the needs of ***(a)*** and ***(b)***.

## 1.3 Challenges

In a research study conducted by Frank et. al. [85], related to social psychology and communications, authors reveal that the accuracy range of detecting deception and misinformation by a normal human being (non-expert) is only 55–58%, which is only slightly better than a guess. In recent years, many fact-verification and news debunking websites such as "*FactCheck.org*" and "*Snopes.com*", etc., have sprung up rapidly, where trained credibility assessment experts and domain experts sit together and try to make consensus about veracity of news. A typical fact or news review process at these organisations[5] involves a very exhaustive Web search; consulting with many of the experts; a thorough review of publications and available evidences. Referring to all such disparate sources and related information provides required context needed to establish the truth-fullness or correctness of news or social media posts. This manual verification of amiss content is a very slow and not a scalable process. To address the scalability issue, automated misinformation detection and fact verification techniques are proposed.

However, the automatic veracity prediction is a very challenging task, and constitutes of several underlying challenges. *First*, finding the **evidences** for a news item or social media post is not a trivial task. It requires

[5] `https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/`

great time and effort to curate and annotate a labelled dataset comprising of multiple evidential records, originating from various sources, corresponding to news items and social media posts. *Second*, modelling the **adequate context** for the news piece is very crucial for the detection accuracy, as mostly only a textual description of a news or social media post lacks required cues pertaining to it's verity. *Third*, the **black box** nature of the machine learning oriented amiss content detection methods poses a serious challenge in terms of understanding the decisions made by the methods. Insights related to inner working of the model can provide significant indications about the reasoning performed by the model for a particular case. *Fourth*, the **early detection** of the amiss content is highly desirable, as we want to curb it's spread as soon as possible. In early stages however, the social media post or news may not have sufficient features and user interactions, for a detection method to perform well.

## 1.4 Research Questions

The research objectives and research questions (RQs) of this thesis are identified and set by reckoning with the challenges figured in Sec 1.3. Some of the RQs have observably a direct link to the outlined challenges and some have an indirect association.

As discussed in Sec 1.3, evidences are very crucial for veracity prediction but it is a very daunting task to congregate evidences for a sufficiently large dataset. We pose a research question ***RQ1*** that whether it is possible to curate and use external evidences for already available public datasets. Further, we want to investigate if such arrangement is indeed beneficial and has potential to subsume the need of manual handcrafting of evidence sources.

> ***RQ1:*** *How effective are the external evidences in addition to labeled dataset?*

The textual content is the indispensable source of context pertaining to the core theme of the news or social media post. The writing-style and vocabulary usage, divulge significant insights about the intent of the writer to deceive. Most of the initial works have successfully utilized the

linguistic cues and stylistic features [37, 26] for amiss content detection. However, natural language comprehension is not a trivial task, largely due to inherent ambiguity of the natural languages. Our models can definitely get benefited by incorporating the other relevant information (side information such as author of web claims, social context, etc.) along with the textual content, which provides additional context. The modelling of the various other attributes and information with the text content needs to be readdressed as simple concatenation of additional features is not very effective. Hence, with an objective to devise a new method to model the side information, we raise research question ***RQ2***.

> ***RQ2:*** *How to model disparate side information or attributes with text content?*

Research questions ***RQ3*** and ***RQ4***, are both purposed to address the challenge related to the black-box nature of the machine learning models. The misinformation detection task is primarily a classification task, which involves approximation of a mapping function from input variables to output variables. Principally, the classification task produces the probability for each class for an instance, which is a very abstract output and it does not provide any insights related to, How did the model arrive at this decision? Machine learning models tend to learn different biases and prejudices [11] from the training data, such as ethnic bias, which can result in unfair decisions in many application scenarios such as image classification and recommendations. If we can generate explanations for the model's decisions, it would give us a tool to peek inside and observe the inner-working of the model so that we can get rid of these biases.

If we talk about the misinformation detection particularly, explanations may include the extraction of words or sentences from the evidences, based on which the model has decided to refute or accept the credibility of the news or social media post. Further, we can also provide users with a concrete summary of explanatory sentences, extracted from various source documents. The explanations also serve the purpose of user friendliness for a fact-verification or debunking system, deployed at a traditional news media or hosted as a service on the Web, to be used by the users.

> ***RQ3:*** *Can we generate meaningful and coherent evidential summaries for amiss contents?*

> ***RQ4:*** *Can we generate explanations for model's decisions?*

The generative adversarial methods such as Generative adversarial network (GAN) [65], are very popular at present. Specifically, GANs are being utilized in many interesting use-cases such as data augmentation and synthetic text and image generation. Research question ***RQ5*** aims at investigating the suitability and usage of synthetic data generation in case of amiss content detection.

> ***RQ5:*** *Can we use generative adversarial methods to improve the model performance?*

It is important to detect and curb all of the amiss content circulating on the Web but it's futile to detect a stale news, which has already spread all around. It is very prudent and productive to detect and check the misinformation as soon as possible so that it does not reach many people [23]. Nonetheless, it is overly strenuous to predict the veracity in early stages of the news propagation. In particular, misinformation detection methods for social media platforms perform poorly in early stage of news as they utilize various social network features and temporal characteristics such as user stance, user interaction and response etc. [27], which are often absent in early stages. Thus, we formulate a research question ***RQ6***, which deals with the challenges related to early detection of amiss content.

> ***RQ6:*** *How early can we detect the misinformation propagation on social media?*

Patterns of user interactions with news item provide significant signals regarding it's credibility [29]. Specifically, in case of proliferation of rumors and misinformation on social media platforms, we need more contextual cues and signals than just simple textual content. Features like

topology and dynamics of the propagation, social network attributes of the users involved in the news propagation, influence relationships among the users are very useful, if modelled in addition to text. Affinity and influence relationships are crucial for information propagation [74]. By answering the research question ***RQ7:***, we envision to devise a method to model the influence relationships among the users, which would provide us with better contextual information related news propagation on social media platforms. In addition, modelling affinity relationship can also reveal, which influencer or high prestige user is mostly responsible for the propagation. In essence, the research question ***RQ7:*** not only acts towards serving the "adequate context" challenge, discussed in Sec 1.3 but also it deals with "black-box" nature of the models by providing clear insights from the learned model.

> ***RQ7:*** *How to model the behaviour of the users and communities with respect to spread of fake news?*

Research question ***RQ8*** is directed towards a specific challenge related to modelling the context for significant numerical values present in news headlines or social media posts. These cardinal values can be currency amount, counts of people, months, years or objects, etc. For an example, consider a news headline *"£100 to play truant! Schools accused of bribing worst pupils to stay away when Ofsted inspectors call"*[6], which holds a contextually important figure i.e. *"£100"*. Most of the prior works in case of deception detection, fail to capture context pertaining to the cardinal values. By asking the research question ***RQ8***, we want to investigate the importance of cardinal values in case of deception or clickbait detection and propose a potential solution for the same.

> ***RQ8:*** *How to capture context related to important cardinal values?*

[6] `https://www.dailymail.co.uk/news/article-2082885/Schools-accused-bribing-worst-pupils-stay-away-Ofsted-inspectors-call.html`

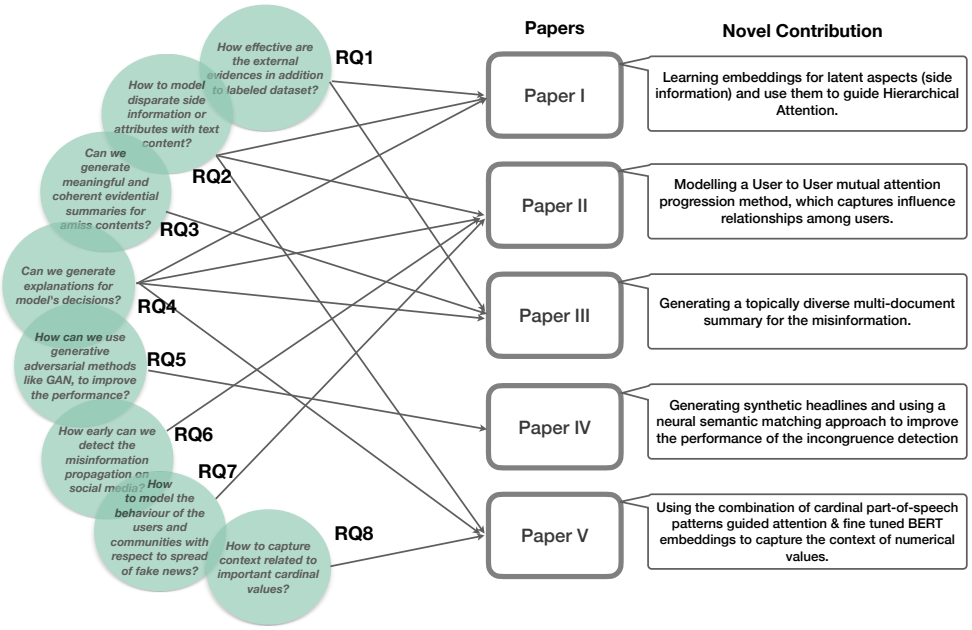



Figure 1.4: Mapping research questions with the papers.

## 1.5 Main Contributions

In this section, we present the main contributions of this thesis. In nutshell, the major contributions in terms of 5 research papers and their relationships (as depicted in Fig 1.5) with the 8 identified research questions as discussed in sec 1.4 are outlined as follows:

**C1. We propose a method to model the latent aspects (side information) with the news text content and extract evidence snippets for interpretability. (Paper I):**

We introduce a novel approach to jointly learn the latent aspects of the news using hierarchical attention mechanism in presence of external evidences extracted from the Web *(caters to research questions **RQ1** and **RQ2**)*. These learned latent aspects provide auxiliary context associated with the news items apart from the text content, which helps in veracity prediction. We also extract evidential text snippets from the external evidences for the sake of interpretability and transparency of the model *(caters to research*

*question **RQ4**)*. The effort in this paper, also resulted in a patent[7] **(US Patent 10,803,387)** and a startup (https://www.factiverse.no).

**C2. We devise a mechanism to model the influence relationships among the users. (Paper II):**

We invent a mechanism to learn direct influence and affinity relationships among the users, present in propagation path of the news on social media *(caters to research questions **RQ2** and **RQ7**)*. Further, we propose an extension of the proposed model to capture the indirect influence relationships. We learn the user embeddings using follower and re-tweet networks via network representation learning methods. Considering the interpretability of the model, we envision to visualize the attention maps to gain some insights concerning the importance of influencers in news dissemination on social media *(caters to research question **RQ4**)*. The proposed model also outperforms the other methods on early detection performance by significant margin *(caters to research question **RQ6**)*.

**C3. We present an approach to generate topically diverse multi-document summaries for the misinformation. (Paper III):**

We propose a model, which incorporates a claim text and a title text driven hierarchical attention by utilizing external evidences to predict the veracity of the Web claims*(caters to research questions **RQ1**)*. Furthermore, we present an algorithm to generate topically diverse, multi-document, and explainable extractive summaries of the evidences for the misinformation *(caters to research questions **RQ3** and **RQ4**)*. We also release a test-collection for misinformation detection, pertaining to climate change and health care.

**C4. We investigate the usability of synthetic headline generation and propose a semantic matching based solution for news headline incongruence detection. (Paper IV):**

We explore the generative adversarial methods such as GANs, to

[7] `https://patents.google.com/patent/US10803387B1/en`

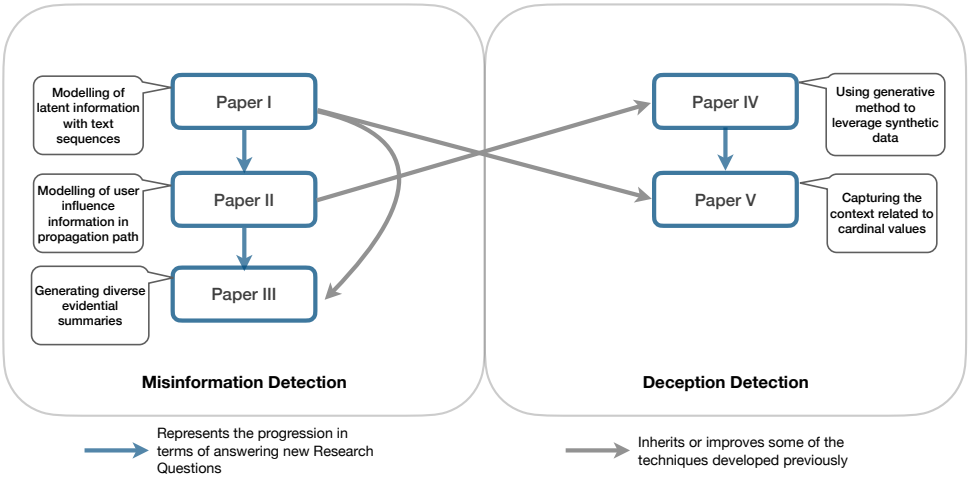



Figure 1.5: A thumbnail illustration of the relationship between contributions.

generate synthetic headlines using the news body content and propose a deep mutual attention based semantic matching to detect the in-congruence with the original news headline *(caters to research questions **RQ5**)*. This proposed architecture resolves several shortcomings and limitations of prior works. We suggest two other variants of the model and a clubbed model, which outperforms all the individual models.

**C5. We define the task of news headline incongruence detection in presence of cardinal values and propose a solution. (Paper V):**

We introduce a task of news headline in-congruence detection in presence of the significant numerical values in the headline. We propose a baseline and a solution, which employs a part-of-speech patterns guided attention mechanism to capture the context specifically related to cardinal values *(caters to research questions **RQ4**) and **RQ8**)*. To showcase the efficacy of the proposed scheme, we conduct an ablation study and extract the attention maps *(caters to research questions **RQ4**)*. We release the derived version of the dataset in which all the news headlines contain cardinal values.

## 1.6 Origins

There are $5$ research papers listed and included in the thesis, out of which $4$ papers are published already in international conference proceedings and the $5^{th}$ paper has been accepted and is currently in publication.

### 1.6.1 Papers

**Paper I** Rahul Mishra ( UiS, Norway ) and Vinay Setty ( UiS, Norway ). **SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection**, in proceedings of the 9th ACM SIGIR International Conference on the Theory of Information Retrieval ( ICTIR '19 ), October 2–5, 2019, Santa Clara, CA, USA.
[ *Associated with* ***RQ1, RQ2, RQ4*** *and* ***C1*** ]

**Paper II** Rahul Mishra ( UiS, Norway ). **Fake News Detection using Higher-order User to User Mutual-attention Progression in Propagation Paths**, in proceedings of the 2020 The 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops ( CVPRW ), Seattle, Washington, USA.
[ *Associated with* ***RQ2, RQ4, RQ6, RQ7*** *and* ***C2*** ]

**Paper III** Rahul Mishra ( UiS, Norway ) , Dhruv Gupta ( MPI, Germany ) and Markus Leippold ( UZH, Switzerland ). **Generating Fact Checking Summaries for Web Claims**, in proceedings of the the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text. Punta Cana, Dominican Republic.
[ *Associated with* ***RQ1, RQ3, RQ4*** *and* ***C3*** ]

**Paper IV** Rahul Mishra ( UiS, Norway ), Piyush Yadav ( Lero, NUI, Ireland ), Remi Calizzano ( DFKI, Germany ) and Markus Leippold ( UZH, Switzerland ). **MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching**, in the proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, Florida, USA.
[ *Associated with* ***RQ5*** *and* ***C4*** ]

**Paper V** Rahul Mishra ( UiS, Norway ) and Shuo Zhang ( Bloomberg,

United Kingdom ). **POSHAN: Cardinal POS Pattern Guided Attention for News Headline Incongruence**, Accepted ( in publication ) in the 30th ACM International Conference on Information and Knowledge Management ( CIKM 2021 ), Queensland, Australia. [ *Associated with **RQ2**, **RQ4**, **RQ8** and **C5*** ]

### 1.6.2 Patents

**Patent I** Vinay Setty ( UiS, Norway ), Rahul Mishra ( UiS, Norway ). **Deep neural architectures for detecting false claims.**, 2020, US Patent 10,803,38.

## 1.7 Thesis Outline

The rest of this thesis is arranged as follows: In Chapter 2, we discuss the background and research gaps in related works pertaining to amiss content detection. Chapter 3 outlines the research contributions made in this dissertation, with adherence to the limitations and gaps identified in Chapter 2. Chapter 4 delineates the overall framework of the whole amiss content detection system and discusses possible use-case scenarios as road-map for real world application. Chapter 5 contains the key findings obtained in the dissertation, possible limitations and future research directions. Finally, all 5 papers have been attached with some realignment and changes to fit the required format of the thesis.

# Chapter 2

# Background and Related Work

*That which is not, shall never be; that which is, shall never cease to be.*
*To the wise, these truths are self-evident.*

*Lord Krishna*
*Srimad Bhagavad Gita*
*Chapter 2, Verse 16*

Most people in the contemporary world rely on the Web to satisfy their conscious or unconscious information needs. However, blind belief on the information retrieved from the Web can be dicey [80] and can have detrimental effects on financial, political and social spheres. The research community has acknowledged the necessity of contrivances and tools, which can prevail the trustworthiness and factual purity on the Web. To this end, researchers have proposed various automated amiss content detection methods in literature. In this chapter, first we provide pointers to some important recent works that analyze and study amiss contents online, which are related to social science domains such as politics, journalism, and psychology, etc. Further, we discuss an overview of automated misinformation and deception detection tasks, related works and trace their limitations. We also discuss the suitability and reasons behind the superior performance of deep neural attention based models as compared to other standard models.

## 2.1 Amiss Content as a Social Science Problem

There has been an increased interest in the social science research community toward analyzing, identifying and combating the spread of misinformation/deception online. The social scientists have studied and analyzed various factors and aspects related to amiss content such as behavioral, political, economic and psychological factors [33, 8, 13]. Researchers conducted experimental studies with people and observed that **confirmation bias** and **motivated reasoning** are the driving forces behind the recklessness shown by people towards misinformation or deceptive material because of their strong prior beliefs or political orientation [47]. The cognitive psychologists suggest two classes of motivations, which drive people to consume any information; namely **Accuracy-driven** and **Goal-driven motivations** [1]. People with goal-driven motivations eagerly satisfy their information needs without caring much about the authenticity or factual correctness of the information, while those with accuracy-driven motivations focus on the correctness of the information [4].

**Prebunking or Inoculation intervention theory** [36] for countering amiss content suggests two steps to counter amiss content; **First**, by giving preemptive warning to the people regarding the potential political or illicit motivations behind the spread of misinformation. **Second**, a rebuttal of an expected argument that exposes an impending fallacy [35]. The prebunking or inoculation theory has been very successful in real world scenarios [14], such as *award-winning Bad News game* in which users can play a browser based game of generating misinformation (post short text) and try to entice a lot of followers by creating trust. Inoculation messages are sent from time to time to make them aware of potential misinformation [5].

**Digital media literacy interventions** [6, 9, 22] are simple suggestions and tips on how to spot amiss content to the users (such as *"beware of sensational headlines"*), which proves to be very effective in increasing the discernment between fake and true news. The inoculation-based techniques are usually domain specific, therefore they do not scale to a very diverse real world scenarios. On the contrary, digital media literacy inter-

ventions are simple rules in form of quick tips to use digital media, which are more generic and scalable. The three major **drawbacks** of digital media literacy interventions are: **1)** It does not alter the belief of the user (due to political bias). **2)** The impact of digital media literacy interventions tends to diminish over time. **3)** These tips can also have negative effects on genuine news consumption, such as the tip like *"beware of sensational headlines"* can affect the accuracy of genuine but sensational headlines.

**Journalistic interventions** [34] such as Fact-checking websites have brought about unprecedented changes and improvements in awareness of the factual correctness of digital information. Fact-checking websites such as "factcheck.org" and politifact.com are very popular. The major objectives of fact-checking are: **First**, educating web users about potential factors related to misinformation and dissemination. **Second**, encouraging factual clarity in political speeches and campaigns [59]. **Third**, reforming journalistic practices to accommodate a greater inclination towards the pursuit of truth.

**Scope of this dissertation for amiss content as a social science problem** is very limited by design and the main research focus is confined to **"Automatic"** detection of amiss content online by proposing machine learning-based solutions. Furthermore, one of the key objectives and motivations of this dissertation is to automatically detect amiss content online in a domain agnostic fashion (not performing any feature engineering or coming up with handcrafted attributes based on any psychological or journalistic social study performed) by devising specialized machine learning models. The primary target users of the proposed system are the general web users (without any experience in journalistic or media domain experience and capabilities) but the social scientists belonging to domains such as politics, journalism, media communication, etc. can easily utilize the developed system either to get the secondary confirmation regarding their manual findings or to scale their experiments to a larger sample size.

## 2.2 Automatic Misinformation Detection

In the literature, researchers have defined the misinformation based on various factors such as intention behind the news etc., and also they have

proposed many related concepts such as Fake news, Disinformation and Rumor. However, there is no standard and widely accepted definition for these concepts. For sake of simplicity and to avoid ambiguity due to very subtle difference between them, we define the misinformation broadly in very simple terms:

*"Misinformation is the spread of false content online."*

The purpose of misinformation detection is to identify factually incorrect or misleading material on the web, either to thwart before it reaches mass or to remove it from the Web, if it has already spread. There is a host of tasks proposed in the literature, which are either considered as the misinformation detection task or have significant overlap such as Claim Verification [38], Text Entailment [50] or Natural Language Inference (NLI) [56], Stance Detection [72], Question Answering (QA) and Rumor Detection [63, 48] etc.

**Claim Verification:** In a typical claim verification or fact checking task [30], we are provided with a textual claim or fact and a collection of textual sources. The claim needs to be verified as supported or refuted, against the given sources. The claim verification task is the closest matching task to a typical fake news or misinformation detection as it mimics a routine procedure followed by a Web user, who wants to check the credibility of a news item or a social media post by going through multiple news portals and information sources.

**Text Entailment:** Unlike claim verification, in a text entailment or natural language inference task, we are given a pair of assertion/fact and the relevant textual source and we need to predict the label as entailment/contradiction/neutral. Therefore, mostly in NLI tasks, there is a pair of sentences comprising of an assertion statement and a corresponding evidence statement rather than a large collection of sources [76, 30]. The text entailment task can be considered as a sub-task of the misinformation detection. As once we have zeroed in on most suitable evidence source for news or social media post, we can apply text entailment techniques to establish it's credibility.

**Question Answering (QA):** The question answering (QA) task is one of the classical natural language understanding (NLU) benchmark tasks, which may include text passages and question pairs of the reading comprehension type [52]. Some of the question answering datasets may have some additional information such as potential answer options for a question. The QA task looks very similar to the fact verification task but there are few differences: ***First***, questions in QA task contain sufficient context to identify the right answers from the source passage. On the contrary, a fact or claim may require additional information and cues to predict it's veracity. ***Second***, fact-checking or veracity prediction presents more stringent requirement than QA system in a sense that not only we need to find the evidences, which support or refute the claim but also we need to establish if it's factually correct.

**Stance Detection:** The stance detection task is to a great extent identical to the natural language inference or text entailment [15]. Given a target hypothesis statement (rumor or news in case of misinformation detection) and user generated text, we need to predict the opinion or stance of the users regarding the target hypothesis statement. Similar to the text entailment, stance detection is also often used as a sub-task of misinformation detection. Specifically, in case of rumor detection on social media platforms, stance detection is frequently used to collect the attitude of the users towards the social media post, which is a very significant feature for rumor detection models [79, 83].

**Rumor Detection:** In simple words, rumors are spread of false information on social media networks like Twitter and Facebook etc. A typical rumor detection task may involve various sub-tasks such as stance detection and veracity prediction, etc. [54]. In contrast to claim verification task, rumor detection exploits social contexts in form of user engagement and actions [10].

Now we discuss various misinformation detection strategies, their applicability, limitations, and research gaps. We can categorize the misinformation detection methods largely in to two categories, namely: Content-based and Social context based. These broader categories can be further divided into several sub-categories.

### 2.2.1 Content and Style based Methods

The original content of news items and social media posts, be it text or associated images, are the primary and most crucial source of information for misinformation detection. The textual content and the images contained by misinformation are usually deliberately made captivating and arousing to leverage the sentimental vulnerability of the people [40]. content-based methods extract various kinds of features from the news content to capture the underlying intent of the author of the news or social media user. For claim verification or fact-checking form of misinformation detection methods also, textual content forms the basis of factual correctness of the claims. Most of the early proposed solutions [75, 81, 78] for misinformation use Content-based features as these are easily and readily available. The content-based methods typically utilize features like subjectivity lexicons, linguistic cues such as lexical and syntactic information contained in the text.

Many initial works for misinformation detection used a ***lexicon oriented method***, where a handcrafted list of lexicons are maintained to capture the patterns related to truthfulness [16]. These lexicons could include some specific terms, identified as frequently used in fake news, certain verbs and pronouns etc. Lexicons are still getting used in some of the recent works [38]. The ***limitations*** of lexicons oriented methods are: 1) less generalizability, as usually lexicons are too domain and use case specific and can not be used in cross domain use cases; 2) less scalability, as it requires manual labor to create lexicons and automation is not possible. Some ***rule based methods*** are also proposed in the literature [64] for stance detection and fake news detection. These rule based methods use regular expression (RegEx) patterns to extract the features from the text. The ***limitation*** of rule based methods is scalability, as making rules exhaustively for all significant patterns is not possible.

***Lexical and syntactic*** information based methods [71], derive lexical and syntactic features from the news text such as Part-of-Speech (PoS), character n-grams, word n-grams, and word count etc. Character n-grams and word n-grams are contiguous sequence of characters and words respectively, which are usually used to measure the text similarity between two texts. The ***limitations*** of these methods is that these features do not cap-

ture the semantic context and semantic dependencies with in the text [16], which is crucial for the model to understand the core theme of the news.

In recent years, ***content-based deep learning methods*** [66, 41] are introduced for fake news detection and related tasks, which outperform traditional content-based methods with a significant margin. Reasons for better performance of deep learning methods over classical content-based methods are: 1) deep learning methods do not require manual feature extraction and feature engineering and can learn very complex features from the text easily [16]; 2) deep learning techniques learn better context and interdependence among the constituent words of text description of the news. The ***limitations*** of the content oriented deep learning methods are: 1) using only textual content information is not sufficient to learn the all important signals pertaining to veracity of the news or claim *(corresponds to the first and second challenge in Sec 1.3)*, we need to model the auxiliary information *(it forms the basis for research questions **RQ1** and **RQ2**)* such as source reliability, author of the social media post or news item; 2) the black box nature of these models is a roadblock in understanding the model's decisions and inner-working *(corresponds to the third challenge in Sec 1.3)*, we need to extract evidential explanations for the model's decisions *(it forms the basis for research questions **RQ3** and **RQ4**)*.

### 2.2.2 Social Context-based Methods

Social context-based methods are commonly employed in the case of misinformation and rumor detection on social networks such as micro-blogs and discussion forums, as many additional features (side information) are readily available in addition to textual content on these platforms. Researchers have proposed several techniques for misinformation detection on social media, which benefit from various auxiliary information besides the textual content of news, which is collectively called as social context such as user profile, user response/action, social network attributes, temporal pattern features, and propagation path of the news, etc. Most of the early works have presented handcrafted feature engineering-based solutions to use social context [73], which is not scalable and also time-consuming.

The ***user response/action based methods*** make use of textual comments and utterances made by users on a social media or forum post [31, 17]. These methods usually model the stance of the users for the news item using stance detection techniques and aggregate all the stances to conclude it's veracity. In addition to stance of the users, user sentiments are also used in user response/action based methods as sentiments provide additional information regarding the attitude of user towards the news [49]. The major ***limitation*** of the user response/action based methods is poor performance in early detection evaluation. Early detection of the fake news on social media is one of the desirable characteristics of the fake news detection models *(corresponds to the fourth challenge in Sec 1.3)*. User feedback and actions are either missing or have a very limited presence during the initial dissemination phase of news on social networks, which are unable to form intricate context needed for verity prediction *(it forms the basis for research question **RQ6**)*.

The ***interim/temporal pattern-based methods*** [58, 16] leverage the alterations occurring in the attributes of the social media user or the posts in different intervals, these alterations show association with the fake news propagation on social network. The potential ***limitation*** of these methods is the need of heavy feature engineering to extract and compute interim patterns, which is not so scalable and requires domain expertise.

The ***propagation/diffusion path oriented methods*** [3, 62] leverage propagation paths of the news or social media posts on the social networks to identify the misinformation. These works use the "retweets" or "shares" trees/cascades to differentiate between the propagation of fake news and true news. The basic assumption and rationale behind using propagation paths to predict rumors or fake news is that the propagation patterns of fake news will differ significantly from the patterns of real news. The ***limitation*** of diffusion path-oriented methods is that these models are too abstract and do not provide clear insights concerning potential influencers or hubs. In addition, they also do not take into account the affinity and influence relationships between users present in the cascade and miss important context *(corresponds to the second challenge in Sec 1.3)*, which is a very important factor in the dissemination of information on social networks*(it forms the basis for research question **RQ7**)*.

## 2.3 Automatic Deception Detection

The emergence of web technologies have given an unprecedented opportunity to various industries like news/media houses, e-commerce/retail and content oriented industries like micro-blog platforms etc., to develop and leverage large client-base. However, some of these businesses seem to be trying to lure more crowds by resorting to illicit means and methods. Deception is one of such illegal practices, which is used by many content oriented businesses to increase the incoming traffic and clicks. Sensational and catchy lines with factual inconsistencies are written to trick people into reading irrelevant material to drive their hidden agenda. There are many definitions of deception in literature but in very simple terms, we can define deception as:

*"Online deception is the act of writing and spreading the illusive and misleading content with the intention of defrauding users"*

The underlying concept at work behind a successful deceptive content is curiosity or information gap [51]. Curiosity gap is a phenomenon in which people fear that they may miss out on some important information if they skip this news. Deceptive contents exploit this information gap to reap financial or political gains [45]. In this thesis, we focus on a specific kind of deception called as clickbait or news headline incongruence.

**Clickbait or Headline Incongruence:** The clickbaits or incongruent headlines are the most popular form of deceptive contents online. A clickbait news usually flaunts a sensational, appealing and exaggerating headline text, which is not in tandem with the news body text. Users are tempted to click on catchy news stories and are disappointed when they read the news body and find that they have been fooled by the news. Consequently, this discourages them to use the same news portal or source again in the future. In a typical clickbait detection task, we are given a news headline and body pair and we need to predict whether or not they coincide with each other [39, 18].

Several methods and schemes are proposed by researchers in the literature

for detecting deceptive contents. There are three major directions in prior works: 1) Linguistic and stylistic feature based methods. 2) Text similarity oriented methods. 3) User behaviour based methods

The ***linguistic and stylistic feature-based methods*** use handcrafted lexicons, text length, word count and part-of speech and various other lexical and syntactic features to learn a classifier for clickbait prediction [55, 57, 45]. The ***limitation*** of linguistic and stylistic feature based methods is the need of manual feature engineering, which is not scalable and very time consuming. Apart from this, it is required to have some training before one can understand the nuances and significance of these features. Deep learning based methods easily outperform linguistic and stylistic feature based methods, as they can learn more intricate patterns and context without any feature engineering.

The ***text similarity-based methods*** rely on textual content of news headline and news body, to identify the cues related to their semantic similarity or coincidence. Some of the initial works leverage the simple lexical similarity between the headline and the news body, using standard similarity measures such as cosine similarity and jaccard similarity etc. Many recent works utilize deep learning-based methods to learn the congruence between headline and the news body by applying various neural attention mechanisms [18, 44]. There are three primary ***limitations*** of the text similarity based methods: 1) the textual similarity techniques work well with short text pairs but in case of clickbaits or news headline incongruence problem, usually news body contains lengthy textual content. If we can generate a parallel shorter text for the long news body text then the performance of the text similarity based methods can be improved significantly *(it forms the basis for research question **RQ5**)*. 2) The other important challenge is the non-overlapping vocabulary between news headline and the body text, therefore lexical similarity based measure can not be used. We need to model the similarity in semantic space, the potential solution is to use the contextualized pre-trained language models such as BERT [7]. 3) The prior works fail to capture the context related to some specific concepts present in the headline such as presence of an important cardinal value or number *(corresponds to the second challenge in Sec 1.3)*, which can play a significant role in deciding the congruence of the headline *(it forms the basis for research question **RQ8**)*.

Recently, some researchers have proposed ***user behaviour-based methods*** for clickbait detection, which model user tendencies and attention span to clickbait news [42]. User tendencies can be captured via user's interaction with the news like likes, up-vote and comments etc. There are two major ***limitations*** of the user behaviour based methods: 1) User interaction statistics and insights are not available for very fresh news and even many older news articles do not receive any user interaction often. 2) Capturing attention span is very useful for clickbait detection but required hardware and settings to collect user's eye gaze data is possible only in laboratory settings and it is a big bottleneck for a commercial deployment of this model.

## 2.4 Why Deep Neural Attention?

Deep learning has recently enjoyed a resurgence, offering an bewildering array of models with state-of-the-art performance and their applications in various application scenarios such as misinformation detection and prevention. However, deep neural models are often criticized for being a black box. Model explainability is not only a desirable feature for better understanding of model decision, but it is also very important in terms of user-friendliness of natural language processing (NLP) tools for non-expert users. Deep neural attention mechanisms [2] have not only played an important role in current state-of-the-art NLP breakthroughs and helped solve many challenges in sequence modeling, but they have also proven to be an effective tool for gathering insights related to the model's decision-making process.

The sequence modeling is a very crucial sub-field in machine learning, where we model and learn from sequential data such as time series data, speech recognition, and natural language processing (NLP). Modelling and capturing long term dependencies in the sentences or sequences is very crucial for most of the NLP tasks but for the tasks such as fact verification, stance detection and textual entailment, it is ardently needed.

The ***recurrent neural network (RNN)*** models were go-to models for sequence modelling until very recently, but they got replaced by cell state oriented models like long short term memory (LSTM) [87] and gated

recurrent unit (GRU) [60] because of some limitations such as vanishing gradient and exploding gradient. For long text sequences RNN models fail to remember the initial words in the sequence, specifically if we need to model the long term dependencies in the sentences. Because in back-propagation through time, when we calculate the gradients for the initial layers, we use the chain rule by applying a multiplicative equation on the gradients computed for the subsequent layers. If the gradients from the subsequent layers are too small, then the gradients for the initial layers become negligible or they vanish[1] [86].

The cell state based models like ***long short term memory (LSTM)*** use a dedicated memory mechanism called as cell state, which holds the information pertaining to long term dependencies. The information on the cell state is controlled by a special mechanism called as gate such as forget gate. Due to persistent cell state controlled by gates, these models can learn longer sequence of texts easily and perform significantly better than vanilla RNN. However, the vanishing gradient problem still not solved fully, only it's less acute than vanilla RNN[2]. In addition, LSTM also suffers from a very serious limitation of encoding and compressing the entire sequence information into a single representation, which is the final hidden state of the LSTM. This design is not very efficient and scalable, and it limits the LSTM's ability to learn very long sequences.

The ***neural attention mechanism*** offers a reasonable solution for learning very long sequences with LSTM. With neural attention, we don't just rely on the last hidden state of the LSTM, instead we use all the hidden states in forming the overall representation. Using a single layer neural network, we learn weights for each hidden state of LSTM, which is called attention score. The attention weight or score for a hidden state signifies the degree of importance of corresponding word in the sequence for the particular task.

In addition, neural attention also gives us the tools to peek inside the decision-making process of the model. By extracting and visualizing the

[1] `https://en.wikipedia.org/wiki/Vanishing_gradient_problem`

[2] `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

attention weights assigned to individual words or sentences [53], we can discover which patterns are responsible for the particular judgment of the model. Furthermore, neural attention can help to take advantage of ancillary and side-information, in addition to textual content, to create better contextual representations.

# Chapter 3

# Our Tryst with Amiss Content

*It was not born; It will never die, nor once having been, can It cease to be.*
*Unborn, Eternal, Ever-enduring, yet Most Ancient, the Spirit dies not*
*when the body is dead.*

*Lord Krishna*
*Srimad Bhagavad Gita*
*Chapter 2, Verse 20*

The challenges and research questions identified in Chapter 1 form the basis of contributions to this dissertation. In this chapter, we briefly describe the proposed methods and solutions to illustrate how they overcome the mentioned challenges and satisfy research questions. The details presented in this section will be indicative in nature, please refer the respective paper for more details and analysis. At first, we present the contribution in Paper I, which proposes a method to learn latest aspect embeddings for misinformation and coupled with research questions **RQ1**, **RQ2**, and **RQ4**. Next, we introduce a technique for mutual-attention progression in propagation paths in Paper II, which deals with research questions **RQ2**, **RQ4**, **RQ6** and **RQ7**. Subsequently, we describe a scheme to generate fact checking summaries for web claims, which satisfies research questions **RQ1**, **RQ3**, and **RQ4**. Next, we explain a mutual attentive semantic matching for headline incongruence, which is linked to research question **RQ5**. Towards the end, we describe cardinal POS patterns for headline

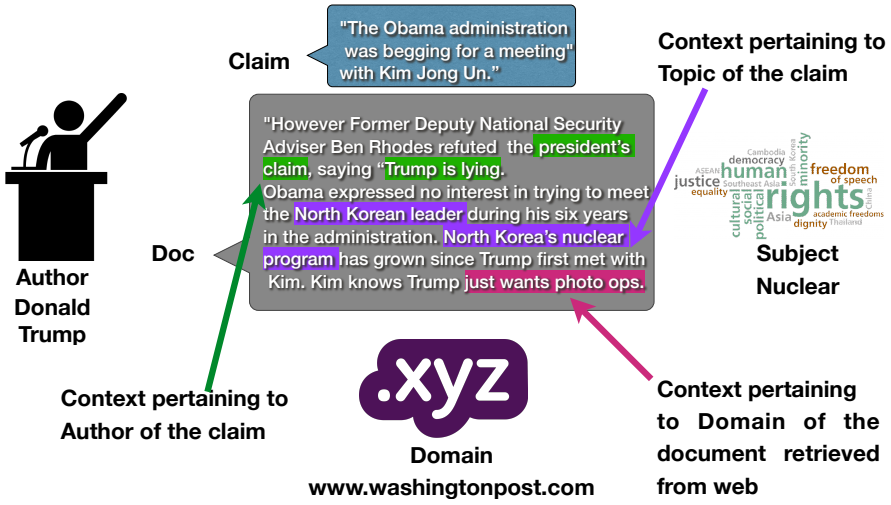



Figure 3.1: Capturing the context related to author, topic and domain in external evidence for the Web claim verification.

incongruence, which answers research questions **RQ2**, **RQ4** and **RQ8**.

## 3.1 Latent Aspect Embeddings for Misinformation (Paper I)

**Objective and Background:** The objective of this work is to predict the veracity of the Web claims. Apart from the textual content of the claim, there are some other attributes also available such as author and subject of the claim. We also utilize textual contents and domain or source information of search results from the Web, which are retrieved using claim text as a query on the Web *( to use them as external evidences as mentioned in outlined challenges in sec 1.3)*. The key objectives of this work are: 1) To use the retrieved search results as external evidences and investigate their usefulness ***(RQ1)***. 2) To leverage the latent or side information such as author, subject and domain to learn better context for the claim ***(RQ2)***. 3) To extract the evidential snippets from the external evidences, which can explain the decisions of the model ***(RQ4)***.

**Method:** We propose a Bi-LSTM [82] based hierarchical attention network, in which attention is guided by the latent aspects; author, topic and domain. By using latent attribute guided attention mechanism *(answers the research question **RQ2**)*, we envisage to learn the intricate contextual cues related to these latent attributes as depicted in fig 3.1. We create three parallel models of hierarchical attention networks; one for each latent attribute and call them author model, subject model and domain model. We concatenate all the three encoded and hierarchically attended document representations, one from each model and apply a softmax classifier on top of this overall document representation, to predict the label. We also propose an algorithm to extract an evidential snippet from the external evidences based on attentions weights learned for words at the word level attention and for sentences at the sentence level attention *(answers the research question **RQ4**)*. Please refer to Paper I for more details.

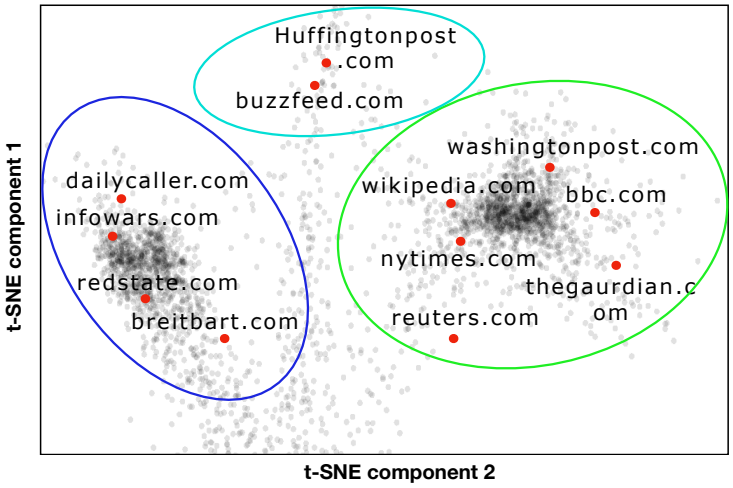



Figure 3.2: Visualization of domain embeddings: We can see clearly separated clusters of trusted and non-trusted domains.

**Discussion:** The proposed model with latent aspect guided attention and external evidence retrieved from the Web outperforms the baselines

and state-of-the-art methods with significant margin for both the publicly available datasets used in evaluation *(answers the research question **RQ1**)*. We also benchmark the performance of the model for textual entailment task, which is a closely related task to claim verification. We visualize the word and sentence level attention weights to investigate the efficacy of the model. Additionally, we extract a snippet from the retrieved document from the web corresponding to the claim. The evidential snippet and attention weight visualization shows that the proposed model captures the useful auxiliary context via latent aspect guided attention. We also visualize the learned embeddings for author, subject and domain attributes and notice that these embeddings learn inherent nuances such as ideology and beliefs of authors in case of author embeddings and trustworthiness in case of domain embeddings as depicted in fig 3.5.

## 3.2 Mutual-attention Progression in Propagation Paths (Paper II)

**Objective and Background:** The purpose of this work is to detect rumors or misinformation on social media platforms like Twitter as quickly as possible *( caters to early detection challenge defined in sec 1.3)*. Apart from the news text, we also have follower network and retweet network ( trees ) available. The propagation or diffusion path of the news in terms of retweet cascades is represented as a variable length multivariate time series of users who retweeted the news. The major objectives of this work are: 1) To propose a model, which performs better in early detection performance compared to the baselines and state-of-the-art ***(RQ6)***. 2) To model the behaviour and influence relationship of the users with respect to misinformation propagation ***(RQ7)***. 3) To utilize the social network information related to users such as follower-following and tweet-retweet relationships along with news text for veracity prediction ***(RQ2)***. 4) To extract insights pertaining to users, who are influencers in the propagation path of misinformation ***(RQ4)***.

**Method:** To begin with, we learn embeddings for all users present in propagation path of the news. We use unsupervised network represen-

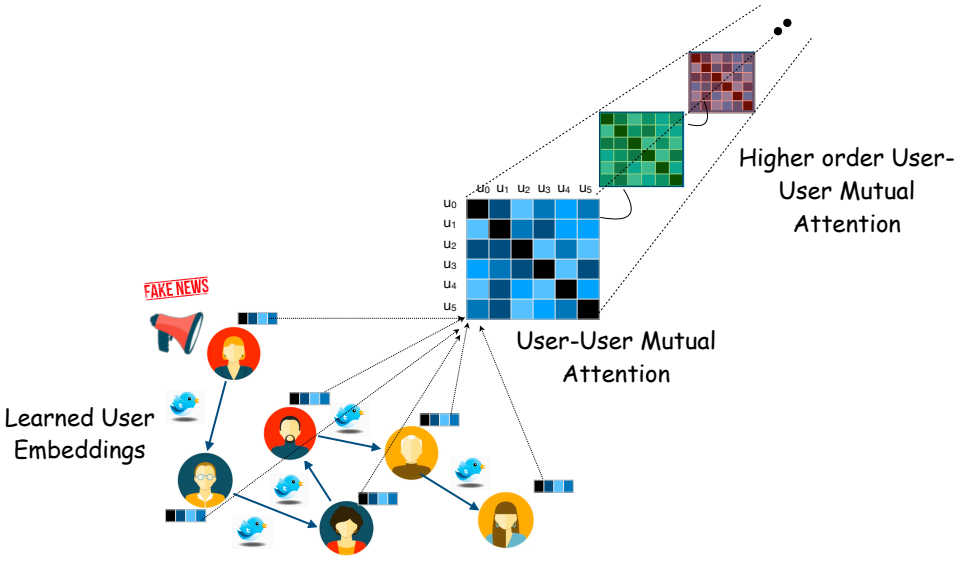



Figure 3.3: Depiction of user-to-user mutual attention and higher order mutual attention computation, using learned user embeddings from retweet and follower networks.

tation learning methods such as DeepWalk [68] and Node2Vec [46], to learn the user embeddings in both follower and retweet network. For each user, we concatenate the corresponding embeddings learned using follower and retweet network, to get the overall embeddings for the users *(answers the research question **RQ2**)*. Next, we use a dense layer to get a score for all possible pairs of the users ( represented in terms of their overall embeddings ) present on the propagation path, which gives us a score matrix. We apply a row wise max pooling on score matrix to get the mutual attention score for each pair of users. We call this mutual attention score as affinity or influence scores or weights *(answers the research question **RQ7**)*. These learned attention weights are multiplied with original sequence of user embeddings to get the attended representation. In addition, we encode the original sequence of user embeddings using a LSTM unit and this encoded sequence is concatenated with attended representation of the sequence to get the final representation. Next, we apply softmax classifier on the final representation to predict the label.

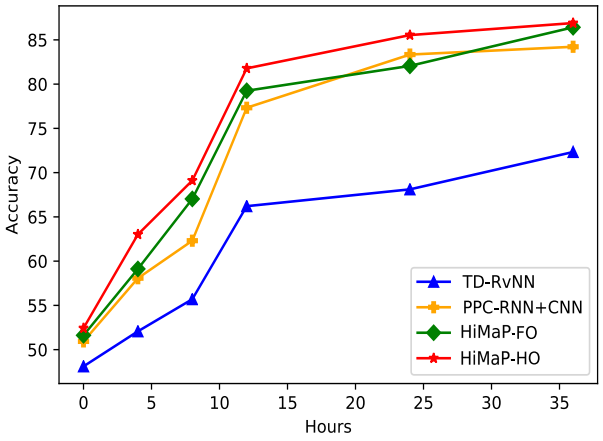



Figure 3.4: Early detection performance comparison with other models

We also introduce an extension to our model, which captures indirect influence relationship among users. For sake of explanability, we visualize mutual attention weights, which reveals which users are influential and have played a significant role in news dissemination *(answers the research question **RQ4**)*. Please refer to Paper II for more details.

**Discussion:** Our model performs well in early detection performance compared to other methods with both the twitter datasets *(answers the research question **RQ6**)*. The proposed higher order mutual attention captures new and uncovered patterns. The visualization of mutual attention scores suggests that users with more number of followers play a key role in misinformation diffusion on the social network. The proposed higher order mutual attention trick can also be useful in various other applications such as mutual attention among the words of a sentence.

## 3.3 Fact Checking Summaries for Web Claims (Paper III)

**Objective and Background:** The objective of this paper is to generate a topically diverse, multi-document, and explainable extractive summary

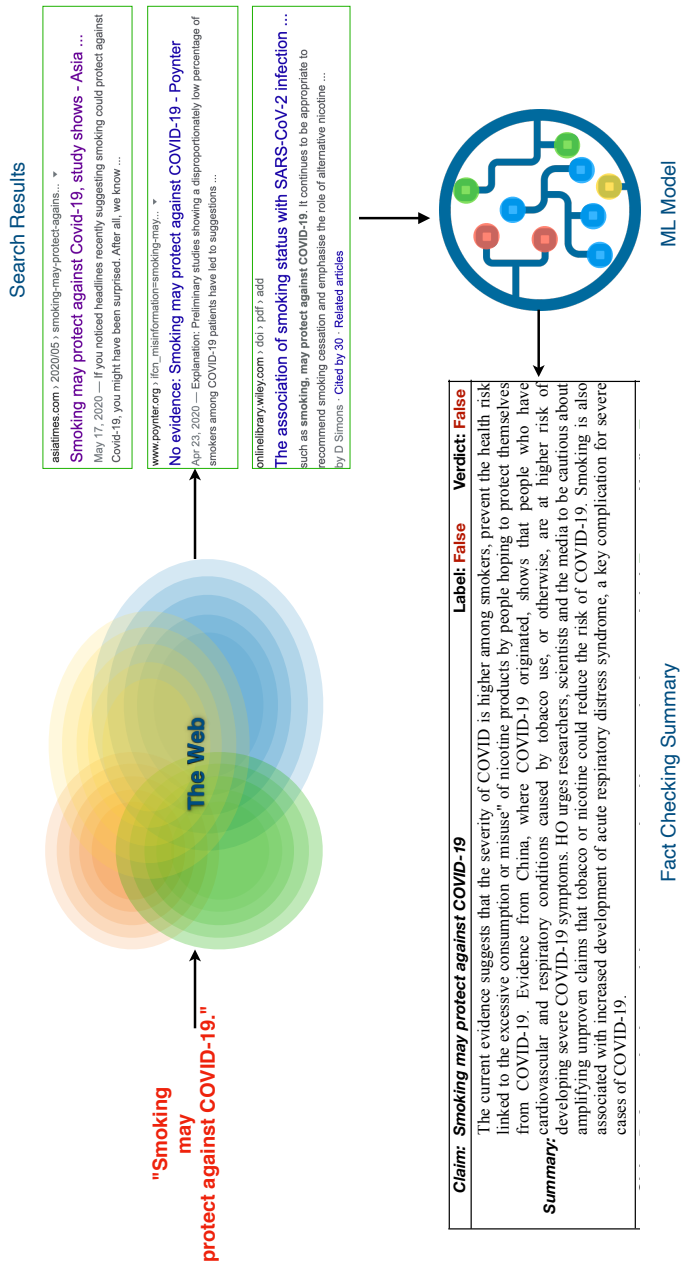



Figure 3.5: Generation of explainable multi-document summary for Web claims

for the misinformation, in addition to predicting the correctness. We have a very similar settings as Paper I; the Web claims to be verified, Web documents retrieved as external evidence using claim text as query, additional attributes such as domains and titles of the retrieved documents, subject and author of the claims, etc. The salient objectives of this work are: 1) To generate an explainable summary of the supporting or refuting evidences for the claims *( caters to black-box challenge defined in sec 1.3)*, which reveals the basis for the model's decisions *(**RQ3** and **RQ4**)*. 2) Improve the usage of claim text (compared to Paper I), to capture better claim related context in external documents *(caters to better context challenge in sec 1.3)*. 3) To utilize the retrieved external documents effectively *(**RQ1**)*.

**Method:** With the objectives of this work in mind, first, we suggest an improvement to Paper I, which simply concatenates the claim text with external document text. We introduce claim driven hierarchical attention to attend salient words and sentences, which are related to the claim . Second, in addition to the claim driven attention, we also use external document's title to guide the hierarchical attention so that we can capture important sections in the document, which are related to title. We also use hierarchical self attention apart from the claim and title driven attention. We calculate the average of the attention scores obtained from all three hierarchical attention mechanisms to obtain an overall attention weight at both the word and sentence levels. Third, we use the overall document representation to predict the label, using a softmax layer. We propose a set cover based algorithm with a diversification objective to generate a topically diverse, multi-document explainable summary of evidences for the web claims using attention weights learned for words and sentences *(answers the research questions **RQ3** and **RQ4**)*. Please refer to Paper III for more details.

**Discussion:** Using external documents retrieved form the Web, we generate a topically diverse multi-document evidential summaries for claims, which fairs well compared to the baselines in terms of ROUGE [84] metrics *(answers the research question **RQ1** as external evidences are indeed helpful)*. The veracity prediction performance of the model is

also good compared to other more complex models and can be attributed to better claim driven attention and additional headline driven attention mechanisms. The multi-document evidential summaries are more user friendly and insightful compared to snippets extracted in Paper I.

## 3.4 Mutual Attentive Semantic Matching for Headline Incongruence (Paper IV)

**Objective and Background:** This work deals with the deception detection on the Web, specifically the objective is to solve news news headline incongruence. We have pairs of news headline/title and corresponding news body content along with manually annotated labels as "Congruent" or "Incongruent" in the datasets. The key objectives of this paper are: 1) To propose a solution for the problem of news body lengthiness, which is a bottleneck for textual and semantic similarity computation. 2) To leverage generative adversarial network based synthetic generation for clickbait detection *(**RQ5**)*. 3) To come up with a semantic matching approach to compute the congruence between news headline and news body.

**Method:** This paper borrows some of the techniques from Paper II; user-to-user mutual attention is adapted to the word-to-word mutual attention in this work, with an expectation learn the complex contextual relationship among the words present in given sequences *(caters to better context challenge defined in sec 1.3)*. The word-to-word mutual attention is computed by utilizing GloVe [67] embedding of the words in very similar fashion as in Paper II, user-to-user mutual attention is computed. Therefore, for the sake of brevity, we are leaving out the same details here. In contrast to Paper II, where user-to-user mutual attention was computed within the users present on propagation path, in this work we compute the word-to-word mutual attention between the pair of words coming from two different sequence of words. The first word sequence is the original headline of the news item and the second word sequence is synthetically generated headline via generative adversarial network using news body text *(answers the research question **RQ5**)*. Please refer to Paper IV for more details.

**Discussion:** The word-to-word mutual attention learns intricate pattern related to semantic matching between the original and synthetic headline, which results in better incongruence detection accuracy, when compared with other methods. Generative adversarial network based synthetic headline generation proves to be critical to the success of this proposed model as it significantly reduces the length of news body content by creating parallel synthetic headlines. We experiment with different headline generation techniques and notice that performance of the model changes significantly with different headline generation technique.

## 3.5 Cardinal POS Patterns for Headline Incongruence (Paper V)

**Objective and Background:** The purpose of this paper is to investigate and solve a specific case in clickbaits or news headline incongreunce task, identified during experiments of Paper IV. We notice that our model in Paper IV and also other models are performing poorly in a particular case, when news headline contains an important cardinal value. We create some new features form the original publicly available datasets such as cardinal phrase and cardinal part-of-speech patterns, rest of the settings are very similar to the Paper IV. The main goals of this paper are: 1) To propose a model, which can perform well in presence of cardinal values in news headlines *(**RQ8**)*. 2) To investigate the usefulness of part-of-speech patterns and cardinal phrases in clickbait detection *(**RQ2**)*. 3) Generate the explanations for the model's decisions *(**RQ4**)*.

**Method:** This paper adapts the techniques developed in Paper I to the clickbait detection settings; we learn the embeddings for part-of-speech patterns in a very similar fashion as we learned the embddings for latent attributes (author, subject and domain) in Paper I. We use cardinal phrase and part-of-speech pattern guided hierarchical attention in addition to the self hierarchical attention to capture the context pertaining to the cardinal values *(answers the research questions **RQ2** and **RQ8**)*. Overall attention score at both the word level and sentence level is calculated by averaging the individual attention scores from all the three attention mechanisms.

We fine tune and use pre-trained language model BERT [7] and extract the word embeddings for the words present in the headline and the body of the news. Please refer to Paper V for more details.

**Discussion:** The proposed model outperforms the other methods as it gives adequate importance to cardinal values present in the news headlines, by incorporating cardinal part-of-speech pattern and cardinal phrase driven attention mechanism. We conduct an ablation study of the model, which reveals the significance of part-of-speech pattern based attention mechanism. We also visualize the attention weights learned using overall attention, which shows the effectiveness of the model and provides evidential insights related to the decisions made by the model *(answers the research question **RQ4**)*.

# Chapter 4

# The Comprehensive Framework

*He who has no love on any side, who when he finds good or evil, neither rejoices nor hates - his wisdom is firmly set.*

*Lord Krishna*
*Srimad Bhagavad Gita*
*Chapter 2, Verse 57*

In this chapter, we define an overall framework of the contributions made to this dissertation, delineating the broad contextual settings and problem description. The objective of this framework is to describe and formalize the amalgamation of all the constituent individual components/contributions with an underlying road-map to perform amiss content detection. We also discuss a user journey mapping/user flow scenario with the help of a flow diagram, to depict and trace the overall process of amiss content detection with respect to the requirement of the target users. In addition, we also discuss potential commercial and social factors and perspectives that we envisage with respect to the real world application of the developed system.

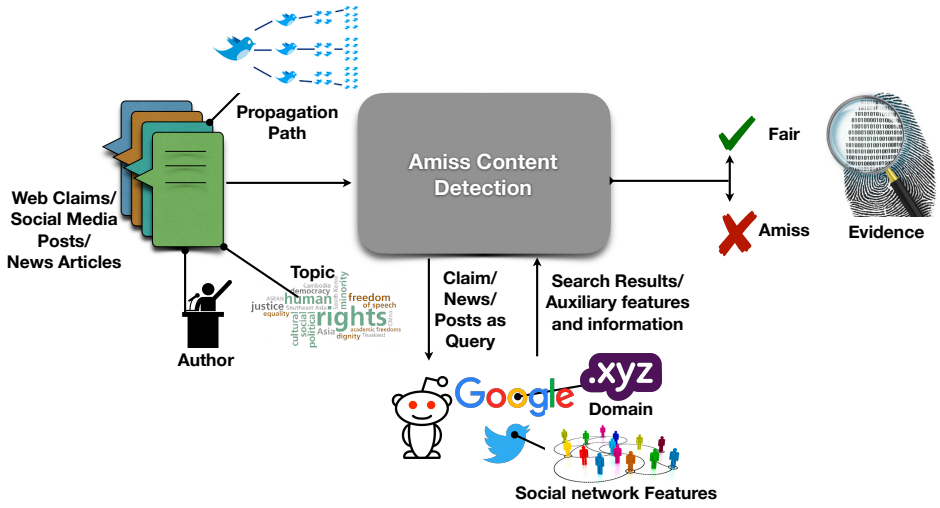



Figure 4.1: The Overall Framework of Amiss Content Detection System

## 4.1 The Overall Framework

We present a high level architecture of amiss content detection system in Fig 4.1. Given a Web claim or news item or social media post in textual form, along with some auxiliary information (side information) such as latent aspects (such as topic, author, domain, etc.), propagation or diffusion path on social media platform and linguistic patterns (such as part-of-speech patterns), we need to asses and predict whether the content is 'Fair' or 'Amiss'. Further, to enhance and capture the contextual information pertaining to the news or social media posts, we also collect and utilize external evidences and features not present in original content (such as retrieval of the textual content of search results or candidate relevant documents from the Web for textual claims and social network features for social media posts). All of these input elements (original textual content, auxiliary information and contextual features) are fed to amiss content detection model.

The amiss content detection model uses novel neural attention-based deep neural networks specially designed to cater to various challenges and

research questions identified in Sec 1.3 and Sec 1.4 respectively. The amiss content detection model not only classifies the news item or social media post as 'Fair' or 'Amiss' but also produces explainable evidential summaries and visualizations for the users to interpret the decisions made by the model.

We now discuss the framework in depth and in a more holistic manner, as depicted in Fig 4.2. We introduce various underlying components or modules as mentioned in the overall pipeline in Fig 4.2. First of all, we explain and present the sources of the bench-marking datasets used in the whole system and argue about their suitability. Next, we describe data preprocessing and transformations, which are necessary to maintain the sanity of the dataset. Furthermore, we present the Representation Learning module, which is used to learn better contextual representations for the data. Next, we briefly talk about the sequence encoders used in the system, which are used to learn and encode text sequences in the form of words and sentences. Subsequently, we present neural attention schemes developed to overcome the challenges and limitations identified in Sec 1.3. Next, we explain the classification module, which produces the outcome of the amiss content detection, for a news piece or social media post as 'Fair' or 'Amiss'. Finally, we discuss about the evaluation and analysis of the results in terms of important parameters such as explainability of the model's decision.

### 4.1.1 Data Collection

As discussed in Sec 2.2, there are many types and flavors of amiss content prevalent online therefore evaluation of amiss content detection model can not be limited to a particular kind of amiss content detection dataset. To this end, we use and evaluate the models with a variety of datasets and settings. All the underlying methods/components developed in this dissertation are benchmarked with publicly available and research community standard datasets for evaluation and performance comparison purposes.

**Web Claim and Entailment Datasets:** We use two publicly available datasets containing political Web claims from the two popular fact-checking websites; Snopes.com and Politifact.com, which are manually

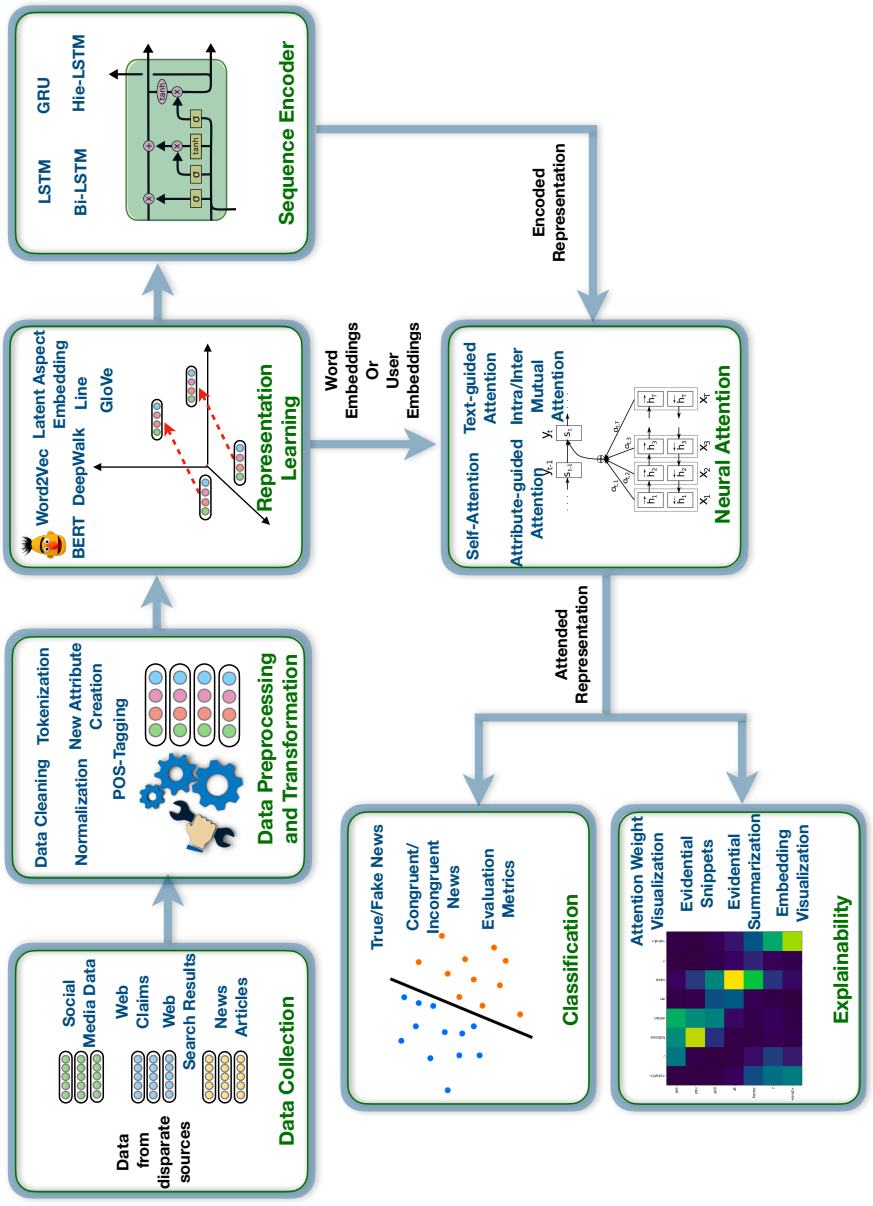



Figure 4.2: A detailed outline of the Overall Framework

handcrafted and annotated by the experts. Popat et al. [24] release these datasets along with corresponding retrieved documents from the Web, using claim text as query. These datasets are very suitable for Web claim veracity prediction task as the presence of Web retrieved documents as external evidence *(related to research question **RQ1**)* gives us opportunity to extract evidential and explainable snippets to support the model's decision *(related to research questions **RQ3** and **RQ4**)*. Textual entailment task is a sub-task of misinformation detection as discussed in Sec 2.2. We use the Fever dataset released by Thorne et al [30], which is considered significant and very suitable bench-mark for the entailment task.

**Social Media Datasets:** Detection of rumors on social media platforms is one of the most important and crucial tasks of detecting amiss content. We use two publicly available Twitter datasets called as Twitter15 and Twitter16, released by Ma et al. [3]. These datasets are particularly important and used for tasks related to early detection of rumors on social media platforms, which is one of the challenges outlined in Sec 1.3 *(related to research question **RQ6**)*. Also, the social media setting enables us to use user behavior and interactions with news or posts to capture relevant signals of veracity and factuality *(related to research question **RQ7**)*.

**News Headline Incongruence Datasets:** News headline incongruence or clickbait is a very prevalent form of amiss content on the Web. We make use of two publicly available benchmark datasets; NELA17 and Click-bait Challenge. The NELA17 dataset is released by Yoon et al. [18] and the Click-bait Challenge dataset is provided by Potthast et al. [25]. We select these two datasets as they come from two disparate sources. The NELA17 dataset is created from the news articles and deliberately curated from reputed and mainstream sources, malignant sources, satire and hyper-partisan sources [20]. On the hand the Click-bait Challenge dataset is created and collected from user generated posts from social media platforms. Consequently, both of these datasets present unique set of challenges for the amiss content detection model, which gives us chance to solve interesting research questions such as **RQ5** and **RQ8**.

### 4.1.2 Data Preprocessing and Manipulation

The purpose of data preprocessing and manipulation is to make sure that the underlying dataset is precise, consistent and meaningful. We make use of various data preprocessing and manipulation steps and techniques with the datasets (as mentioned in the Sec 4.1.1) used in the system such as normalization, tokenization, padding, new feature creation etc. For sake of brevity, here are some of the important data prepossessing and transformation use-cases from our proposed system:

- In case of Politifact and Snopes datasets, we use cosine similarity score-based threshold to include only highly relevant parts of the documents, which are retrieved from the Web as external evidences for the claim *(related to research questions **RQ1** and **RQ2**)*.
- We use Latent Dirichlet Allocation-based topic model to generate topics for the textual claims in Fever dataset as topic information was not present in the original dataset *(related to research question **RQ2**)*.
- We augment the Twitter15 and Twitter16 datasets with follower-following information by crawling the twitter social network for the users involved in re-tweet paths of the news items present in the datasets *(related to research questions **RQ6** and **RQ7**)*.
- We use generative adversarial network (GAN) based synthetic news headline generation methods such as stylistic headline generation (SHG) [28] to produce auxiliary headlines for both the NELA17 and Clickbait Challenge datasets *(related to research question **RQ5**)*.
- We create two new attributes in NELA17 and Clickbait Challenge datasets called as Cardinal Part-of-speech (POS) Pattern and Cardinal Phrase by using Part-of-speech tagging and regular expression-based rules *(related to research question **RQ8**)*.

### 4.1.3 Representation Learning

The adequate representation of the input features plays a very significant role in successfully learning from the training dataset for an underlying task. Representation learning techniques enable us to learn a very low

dimensional and generalizable representation of an input feature, which captures various complex contextual aspects related to the input feature such as the semantics for words. We employ three kinds of semantic representation learning methods in this dissertation to represent the input features namely: Word Embedding, Latent Aspect/attribute Embedding, and Network Representation Learning.

**Word Embedding:** Word embedding is a technique for obtaining very low dimensional dense representations or real-valued vectors in the semantic vector space for words occurring in a text sequence. We utilize and experiment with various word embedding methods in all the contributions made in this dissertation such as Word2Vec [70], GloVe [67], and BERT [7] etc. The Word2Vec and GloVe word embedding techniques are context independent word embedding methods whereas BERT is a context sensitive word embedding method.

**Latent Aspect/attribute Embedding:** Auxiliary information/attributes hold significant hints and pointers relating to the context of the underlying textual content. In experiments, we observe that simple concatenation of these auxiliary/latent information with textual content is not very effective overall representation therefore in this dissertation, we propose to learn contextual embeddings of these latent information to model them effectively with textual content *(related to research question **RQ2**)*. In case of veracity prediction of Web claims, we learn embeddings for latent attributes such as Domain of retrieved document (external evidence for the Web claim) from the Web, Topic of the Web claim and Author of the Web claim, which are trained during training time using hierarchical attention mechanism and utilized during test time. In similar fashion, we learn embeddings for cardinal part-of-speech (POS) patterns for news headline incongruence detection *(related to research question **RQ8**)*.

**Network Representation Learning Methods:** Network representation learning (NRL) methods are very important recent advances in the field of network structure learning and modeling. NRL techniques have surpassed and outperformed handcrafted feature engineering based methods for extracting features from networks such as social media networks, scholarly

citation networks, and biological networks. In case of early rumor/misinformation detection on social media platform, we use and experiment with various state-of-the-art unsupervised network representation learning methods such as DeepWalk [68], Node2Vec [46], Line [61], and APP[43], to learn embeddings for social media users based on both re-tweet and follower-following networks *(related to research question **RQ2** and **RQ7**)*. We observe in the experiments that the Line method performs better than the other node embedding methods for both the twitter datasets.

### 4.1.4 Sequence Encoding/Modelling

As discussed in Sec 2.4 in detail that sequence modelling is one of the pivotal sub-field in machine learning. The cell state based models such as long short term memory (LSTM) and gated recurrent unit (GRU) are the preferred models to encode text sequences. We use the bi-directional versions of both LSTM and GRU, because the bi-directional versions perform better than the uni-directional versions as they learn the context of both forward and backward directions in the text sequence.

**Hierarchical Sequence Encoder:** A better overall representation of a text document can be obtained by incorporating knowledge of document structure in the model architecture. Hierarchical Encoder provides accurate semantic and structural representation of text documents as text documents are inherently hierarchical in nature. The sequence of words is encoded to form the sentence representation and the sequence of the sentences is encoded to form the overall document representation *(related to research question **RQ2**)*.

### 4.1.5 Neural Attention

Deep neural attention-based models have been proven to be very effective in modelling and learning in case of long sequential data such as textual content *(as discussed in Sec 2.4)*. In addition, neural attention-based mechanisms provide a way to increase the explainability of the model. In Sec 4.1.4, we discussed the hierarchical encoder and its suitability for encoding text documents. Similarly , we also use hierarchical neural attention mechanism in addition to the hierarchical encoder. In a typical

text document, not all parts are equally relevant to the underlying task, therefore, it is necessary to determine the relevant and useful sections. Determining useful sections involves modeling the interactions between words at the word level and between sentences at the sentence level, not just their presence in isolation.

The hierarchical neural attention mechanism gives us a tool to model a text document in such a way that only important and useful parts of the document are given importance. The hierarchical neural attention also captures the latent cues hidden in sentence formation and style. In this dissertation, we propose various novel neural attention mechanisms to meet many of the challenges identified in section 1.3 and to address many of the research questions raised in section 1.4. For the sake of brevity and convenience, we classify the neural attention-based mechanisms proposed in this dissertation into three categories as listed here:

- **Latent Aspect-guided Attention:**
  As discussed in Sec 4.1.3, the modeling and use of latent aspects and side information provide important contextual clues in addition to the textual content. We learn embeddings for latent aspects such as domain, subject and author in web claim truthfulness prediction settings using latent-guided hierarchical attention. The purpose of hierarchical latent aspect-guided attention is to select words at the word level and sentences at the sentence level, that are associated and relevant to the latent aspects (e.g. author: Donald Trump; attended words: United States, President, etc.) *(related to research question **RQ2**)*.

  In a very similar fashion, we learn embeddings for part-of-speech (POS) tag patterns (e.g. NN : CD : JJ) related to cardinal values by using hierarchical attention. The objective of the cardinal POS tag pattern-guided attention is to attend or select salient words that are significant and have some connotation with cardinal phrase present in the text. with the cardinal phrase of the headline *(related to research question **RQ8**)*.

- **Textual Content-guided Attention:**
  In predicting the veracity of a Web claim, the claim text itself provides the most important context. It is very helpful for the pre-

dictive model to identify and use the sections of external evidence documents that are relevant to the content of the claim. The claim text-driven hierarchical attention technique selects key words and sentences, which are important and related to the claim text.

Likewise, we use textual titles of documents/articles, which are retrieved from the Web as external evidences, to guide the hierarchical attention to capture the sections in the articles which are more critical and relevant for the title and overall theme of the article *(related to research question **RQ2**)*.

- **Deep Mutual Attention:**
  It is important to model intra-relationships within sequential data such as text sequences, time series data, etc. For example, in the case of early detection of rumors on social media, it would be very useful to model the influence relationships among users who are on the propagation path of a news item. To model such influence/affinity intra-propagation path relationships, we propose user-to-user mutual attention method, which uses user embeddings learned using unsupervised network representation learning methods (as discussed in Sec 4.1.3) for all the users present on propagation path to compute mutual attention scores. We also propose and use a higher order mutual attention mechanism, which learns multi-hop relationships among the users. In multi-hop relationships influence depends on a group of users rather than a single user *(related to research question **RQ7**)*.

  We use a very similar approach in the case of the news headline incompatibility/incongruence task, in which we have a news body and an associated news title and the objective is to determine whether the title matches or coincides with the news body content. We apply a word-to-word mutual attention mechanism similar to user-to-user attention to model the congruence between news headline and news body content. The key difference in the design of these two mutual attention techniques is that the user-to-user mutual attention is intra-sequence (within a propagation path) whereas word-to-word mutual attention is inter-sequence (between two word sequences; one for news headline and another for news body). The other major differ-

ence between them is the way we compute mutual attention score; in case of user-user mutual attention, we concatenate the pairs of learned user embeddings for attention weight computation whereas in word-to-word mutual attention, we use difference between the pairs of word embeddings.

### 4.1.6 Classification

After getting overall encoded and attended representation, we use softmax classifier to classify the content as 'Fair' or 'Amiss' at coarse granularity level task of amiss content detection, while it can be classified as True/False or Congruent/Incongruent in finer granularity level tasks of misinformation/rumor detection and clickbait detection.

We evaluate the classification results of our system by comparing it with various suitable baseline and state-of-the-art models using various evaluation metrics such as accuracy, F1 score, and area under the ROC curve (AUC) on publicly available and community standard benchmark dataset (discussed in Sec 4.1.1). We also evaluate results for some sub-tasks i.e. summarization of external evidences for a Web claim using appropriate metrics such as ROUGE-1, ROUGE-2, and ROUGE-L scores. The results of the proposed system are also tested for statistical significance using a pairwise Student's t-test. Evaluation performed using publicly available and community standard benchmark datasets is a standard and trusted practice in the machine learning (ML) and natural language processing (NLP) research communities, which is also practiced and encouraged by leading researchers.

### 4.1.7 Explainability and Analysis

We conduct a number of experiments to analyze the interpretability/explainability and effectiveness of the model by extracting and visualizing complex insights. As discussed in Sec 2.4, model explainability has become ardent need for analyzing and understanding the model decisions and neural attention mechanisms enable us to enhance and introspect the model interpretability. We perform various insightful analysis to reveal the efficacy of the proposed system, some of them are listed here:

- **Attention Weight Visualization:**
  We extract attended words and sentences with corresponding attention weights for an anecdotal example from the publicly available dataset. We assign a particular color to each kind of hierarchical attention mechanism (such as domain-driven, author-driven, and claim text-driven attention etc.). The depth of the colors represents the distribution of attention weights. By analyzing the attention scores for different words and sentences in an example, we can infer the reasoning and patterns related to model decisions *(related to research question **RQ4**)*.

- **Embedding Visualization:**
  As discussed in Sec 4.1.3, we learn various latent aspect embeddings (i.e. domain, topic, author, and pos pattern embeddings) using hierarchical latent aspect-guided attention mechanisms. We use t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize these latent aspect embeddings in low dimensional space, which reveals and showcases the effectiveness of the proposed aspect-guided attention mechanism for learning intricate contextual signals and cues.

- **Evidential Summary:**
  Different news providers and social media platform users can reveal different aspects of the same news with different levels of depth, granularity and temporality. Hence, it is very user friendly is to amalgamate multiple sources of information and present a holistic, coherent, and non-redundant evidential summary for the Web claims. We propose a summarization algorithm, which uses attention scores for words/sentences and topical information for each sentence to generate a ranked list of sentences that are: novel, non-redundant, and diverse across the topics identified from the text of the documents *(related to research questions **RQ3** and **RQ4**)*.

- **Early Detection Analysis:** As discussed in the challenges mentioned in section challenges, early detection of misinformation is the need of the hour. We compare the performance of the proposed system with state-of-the-art methods for early detection by plotting the overall accuracy vs elapsed time since the original tweet is

posted. Our model outperforms all the baselines and state-of-the-art methods with a significant margin *(related to research question **RQ6**)*.

## 4.2 The User Journey Mapping

The target users or consumers of research contributions to this dissertation are general web users (including social media users) and deployment platforms are search engine companies (Google, Bing, etc.), social media platforms (Facebook, Twitter, etc.). As discussed in Sec 2.1 the key objective of this dissertation is to automatically identify misinformation in a domain agnostic manner ( not relying on any psychological or journalistic social study performed). We design the proposed system with the assumption that the target users (general web users) do not have any background and experience with any news/social media analytics domain capability.

In real world scenario, if a specific web user wants to determine whether a particular online news/post is 'amiss' or 'fair', the road-map to achieve this may include: issuing a query to a search engine using the text of the news; going through the top results given by the search engines; and finally making an informed decision based on the background information gathered. In a broader sense, with our proposed system, we try to mimic this scenario using artificial intelligence contrivances and help web users to make informed decisions.

In Fig 4.3 , a ***user flow diagram*** for a typical web user is depicted using the proposed overall framework. The user starts by providing a news item/web claim/social media post as input to the system. Next, the system enters the auxiliary data collection phase, where search results from the web and social network features from social networks can be collected. After this, data preprocessing and data transformation steps are performed to convert the data into the required format. In addition, appropriate embedding methods are used to represent the data in vector space (word embeddings, latent aspect embeddings and user embeddings, etc.) before encoding it through the sequence encoder. Next, the system prompts the user for his input whether the user wants to check for clickbait or misinformation. Based on the user's answer, the system uses pre-trained models of clickbait

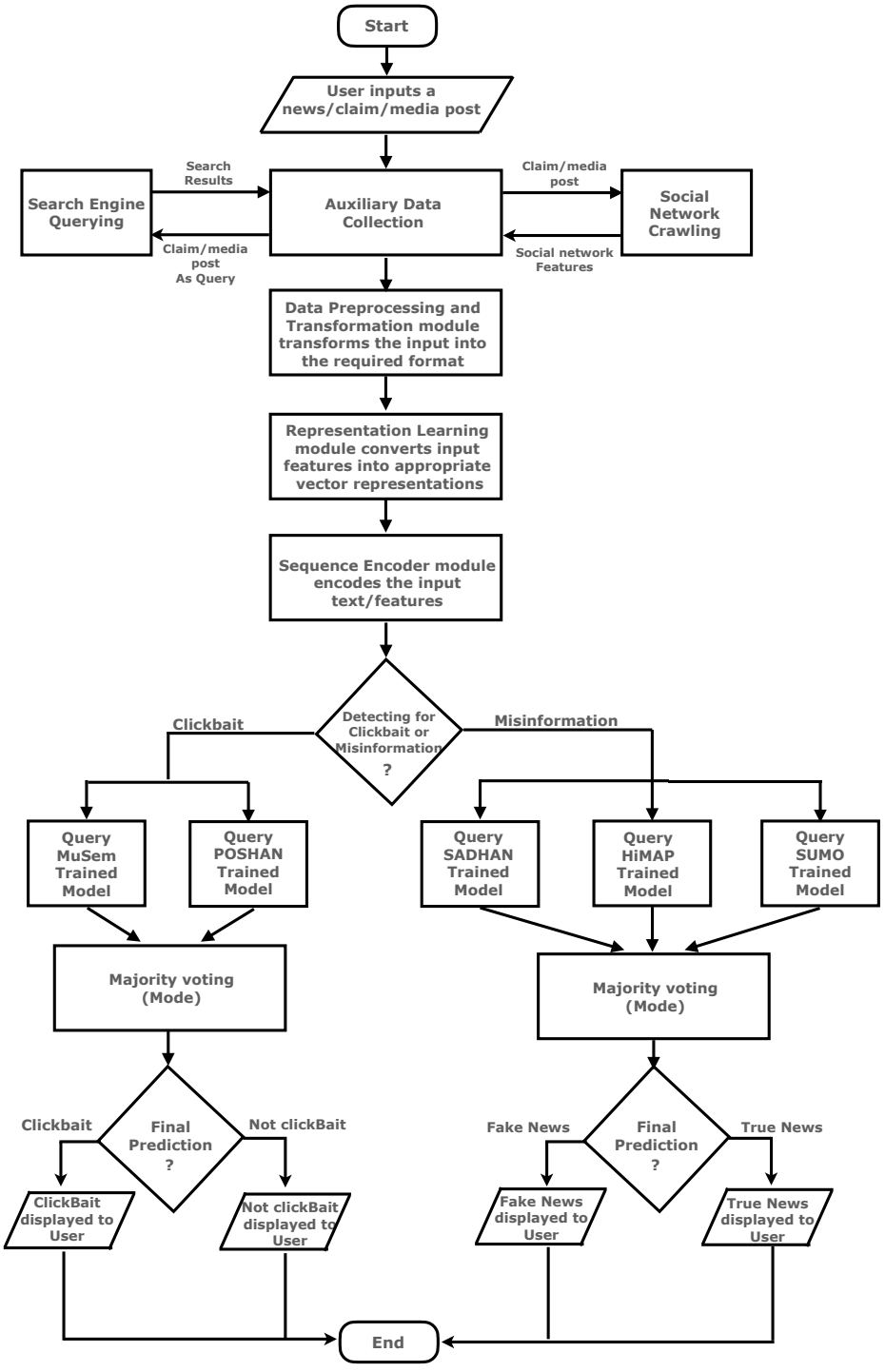



Figure 4.3: A user flow diagram for the overall framework

detection or misinformation detection. Majority voting mechanism is used to aggregate the prediction results from multiple models. Finally, the overall result is shown to the users.

## 4.3 Potential scenarios for real world deployments

- **Browser Plug-in:**
  The proposed system can be incorporated with browsers as a browser plug-in, which is simplest and cost effective deployment of the system. It is also very user friendly and can support all sort of web-browsers including their mobile versions.
- **Integration with Search Engines:**
  Search engines are one of the most adequate options for deployment of the proposed system as a typical web user uses search engines to find out about the veracity of the news. When searching for a news item, the search engine will collect information in the form of external evidence from multiple sources from the web and infer its veracity and a sufficient evidence summary will be displayed to the user.
- **Integration with Social Media Platforms:**
  Another very interesting application scenario would be integration to the popular social media platforms like Twitter and Facebook, as these platforms are prone to fake news and mis-information more readily.

## 4.4 Stakeholders

In any deployment scenario, end-users are one of the most important stakeholders as the overall success of the system depends on the acceptability of the system by the end-users. User-friendliness and user awareness are important factors relating to end-users. Since the normal Web users are the targeted consumers of the proposed system, they are the major stakeholders. There is good scope for integration of the system with popular

search engines such as Google and Bing. Social media, on the other hand, has become increasingly important for such a system. The integration of our system with social media platforms such as Twitter and Facebook can be of great help to end users as these platforms are more prone to fake news and misinformation than traditional news media. Lastly, the government is an important stakeholder as a policy maker for information access and IT legislation.

**Stakeholders Factors:** Based on the various stakeholders identified, we can define various stakeholder factors that are relevant to our system as mentioned below:

- **User Awareness:**
  At a time when misinformation is a widespread phenomenon, it is important to make end-users and other stakeholders aware of the effectiveness and necessity of such a system.

- **Willingness to Accept the Need**
  All the stakeholders should understand the importance and usefulness of such a system.

- **Human Intervention required or not:**
  As the system relies on Artificial Intelligence to generate the adequate results, it could be debatable topic whether to use human in loop approaches with the proposed system due to ethical concerns.

- **Political Correctness:**
  As the veracity decision and analysis can be prone to hyper-partisan and biases, we need to have a mechanism to counter the same.

- **Intellectual Property Rights (IPR):**
  Since the system uses information and articles from various online platforms, blogs and portals, we need to incorporate the collected data by following the policies and norms of governments and content owners. We can filter out content that may result in infringement.

## 4.5 Societal Impacts

There are both positive and negative societal impacts, which are mentioned below.

- **Easier to get Trusted and Crisp News**
  With the proposed system, it would be easier to access the trustworthy content with crisp evidential summaries about the veracity. Early detection of misinformation using user profiling technique is the key differentiator of the system.
- **User Friendly**
  The users will not have to search plethora of websites to get gist of an event story, the system provides users with multi aspects of the news with evidences on the fly.
- **Privacy Concerns**
  User profiling and learning misinformation dissemination patterns may raise some eyebrows due to privacy concerns, but this can be avoided by obfuscating the data as much as possible.
- **Changes in Preferences, Trust and Sentiments**
  As amiss content detection reveals a source of information to be deceitful/fraudulent. This can bring down the reputation of some news portals, bloggers and media houses.

## 4.6 Governance

**User Data Confidentiality and Anonymity:** The deployed system uses only anonymized data from news portals and social media platforms. All personal information relating to the ordinary Web users is appropriately obscured (such as follower-following information on social media). As far as the publication of the results is concerned, all insights generated are privacy-preserving with respect to users and reflect the performance of our model only.

# Chapter 5

# Conclusions, Limitations and Prospects

*He whose mind is not perturbed in pain, who has no longing for pleasures, who is free from desire, fear and anger - he is called a sage of firm wisdom.*

*Lord Krishna*
*Srimad Bhagavad Gita*
*Chapter 2, Verse 56*

The purpose of this dissertation is to detect online misinformation and misleading contents. To this end, we identify salient challenges and research questions and we suggest and experiment with various deep neural attention based models, which are designed to meet the identified challenges and research questions.

## 5.1 Conclusions and Takeaways

The major highlights and takeaways of this dissertation are as follows:

- We propose a method to learn embeddings for latent attributes/aspects

of the Web claims, which are trained at the train-time and used at the test-time to guide the attention to select salient words and sentences relevant for the verity of the claim. The visualization of these latent aspect embeddings reveals that not only cues related to involvement of these attributes (such as author of the claim) in misinformation but also intricate nuances of other concepts such as ideology of authors are also learned by these embeddings.

- The external evidences are quite helpful in misinformation detection. We extract evidential snippets from external evidences, which complements the interpretability of the model.

- We introduce a technique to model influence and affinity relationship among the social media users, which captures the complex patterns pertaining to diffusion of the news on social network. Our model does not rely on temporal network patterns and user responses such as replies and comments, therefore it performs well in early detection accuracy compared to other methods. Further, we extend the proposed model to capture the indirect influence relationships using higher order mutual attention trick. The visualization of the attention maps for a propagation path shows that higher order mutual attention produces uncovered and novel influence patterns.

- We generate topically diverse multi-document explainable summaries for the Web claims, which not only aids in model transparency and interpretability, but can also provide a good user experience in fact-checking plugins and tools for non-expert users.

- We present a synthetic headline generation and attentive semantic matching based headline incongruence detection method. We experiment with various synthetic headline generation methods and notice that the effectiveness of synthetic headline generation step plays a key role in overall performance of the method.

- We define a task of news headline incongruence detection in presence of significant cardinal values in headline. In experiments, we notice that prior works fail to capture context related to cardinal values present in news titles. We present a solution, which uses a novel cardinal part-of-speech pattern driven hierarchical attention

and cardinal phrase driven attention to attend important words and sentences relevant for cardinal conditions and values. An ablation study shows that cardinal part-of-speech pattern driven attention plays a crucial role in overall performance of the model. The visualization of attention maps verifies the significance of the proposed techniques and deciphers the decisions of the model.

## 5.2 Limitations and Implications

We outline the potential limitations and implications of our contributions included in this dissertation along these lines:

(1) The model proposed in Paper I for Web claim verification outperforms other works with significant gains. However, it's very hard to train and requires more hardware and training time due to it's complex structure. In addition, since the model relies on latent aspects for auxiliary context learning, the absence of latent aspects leads to performance degradation. The model also does not include a provision for adding new/unseen values to the latent features at the time of testing.

(2) The user-to-user mutual attention technique proposed in Paper II, suffers from an inherent challenge with long propagation paths. For very long propagation paths with a large number of users present, the computation of pair-wise mutual attention scores becomes computationally expensive and time-consuming. Model performance varies with user embeddings learned using different unsupervised network representation learning methods, so it's important to try and experiment with different network representation learning methods and find the method that performs best for the dataset, which is very time consuming.

(3) The diversification objective in the summarization algorithm in Paper III requires a separate computation of the topic model for each claim. The topic model is applied on all the sentences attended by the veracity prediction model originating from multiple external evidence sources, once for each web claim instance for which we wish to generate an explanatory summary.

(4) The synthetic headline generation-based method proposed in Paper IV uses a low-dimensional representation of the news body text instead of using the original long content, which is very effective. However, by generating a smaller and lower dimensional text for the news body, we may miss out on some important contextual information, such as important cardinal values.

(5) In Paper V, we notice that the part-of-speech (POS) tagger sometimes misses the cardinal values or does a wrong tag assignment, therefore model's performance may vary with different POS taggers.

## 5.3 Future Prospects

There can be many viable future prospects for the contributions made in this thesis as the problem of misinformation and deceptive content is evolving in nature. As a result, methods require novel extensions and features to keep up with changing landscapes and conditions.

- In Paper I, generating explanations for misinformation using attention weights is very useful, but at the same time we need to use additional explainability tools such as learning the disentangled representations [19] to confirm whether they corroborate with each other.
- The user-to-user mutual attention method in Paper II takes sequence of user embeddings learned via follower and retweet networks as input. Therefore, in essence it utilizes social connection and retweet connection information of users but it does not use any heuristics related to user tweet history or liking. This can be very helpful in deciding user behavior towards misinformation and can be used in addition to embedding learned using follower and retweet networks.
- For explainable multi-document summarization of evidential documents for misinformation in Paper III, a user study can be very beneficial in order to get the useful feed-backs and validations.
- In Paper IV, the proposed model for news headline incongruence detection has two sequential parts; synthetic headline generation and mutual attention based semantic matching. We can comp up

with an end to end version of the model, where synthetic headline generation and semantic matching steps are seamlessly integrated.

- The Paper IV borrows user-to-user mutual attention technique from Paper II and adapts it to word-to-word mutual attention. However, we do not use the higher order mutual attention progression trick in Paper IV for words, which can be explored in the future.
- In Paper V, we try to capture context related to cardinal values present in the news headlines. Nonetheless, we do not model or consider the degree of importance of cardinal values.
- In the future, we also look forward to conducting a detailed user study on the informativeness and interpretability of evidence snippets and evidence summaries.

# References

[1] **Kunda Z**. "The case for motivated reasoning." In: *Psychol Bull* 108(3) (1990 Nov), pp. 480–98. DOI: 10.1037/0033-2909.108.3.480.

[2] **Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio**. *Neural Machine Translation by Jointly Learning to Align and Translate*. In ICLR 2015 as oral presentation.

[3] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning." In: ACL'17, pp. 708–717.

[4] **MATHIAS OSMUNDSEN, ALEXANDER BOR, PETER BJERREGAARD VAHLSTRUP, ANJA BECHMANN, and MICHAEL BANG PETERSEN**. "Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter." In: *American Political Science Review* 115.3 (2021), pp. 999–1015. DOI: 10.1017/S0003055421000290.

[5] **J Roozenbeek, R Maertens, W McClanahan, and S van der Linden**. "Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited." In: *Educational and Psychological Measurement* 81(2) (2021), pp. 340–362. DOI: 10.1177/0013164420940378.

[6] **Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar**. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." In: *Proceedings of the National Academy of Sciences* 117.27 (2020), pp. 15536–15545. ISSN: 0027-8424. DOI: 10.1073/pnas.1920498117. eprint: https://www.pnas.org/content/117/27/15536.full.pdf. URL: https://www.pnas.org/content/117/27/15536.

[7] **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Minneapolis, Minnesota: Association

for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[8] **Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer**. "Fake news on Twitter during the 2016 U.S. presidential election." In: *Science* 363.6425 (2019), pp. 374–378. ISSN: 0036-8075. DOI: 10.1126/science.aau2706. eprint: https://science.sciencemag.org/content/363/6425/374.full.pdf. URL: https://science.sciencemag.org/content/363/6425/374.

[9] **S. Mo Jones-Jang, Tara Mortensen, and Jingjing Liu**. "Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't." In: *American Behavioral Scientist* 65.2 (2021/08/24 2019), pp. 371–388. DOI: 10.1177/0002764219869406. URL: https://doi.org/10.1177/0002764219869406.

[10] **Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu**. "Rumor Detection on Social Media: Datasets, Methods and Opportunities." In: *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 66–75. DOI: 10.18653/v1/D19-5008.

[11] **Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan**. "A Survey on Bias and Fairness in Machine Learning." In: *arXiv e-prints*, arXiv:1908.09635 (Aug. 2019), arXiv:1908.09635. arXiv: 1908.09635 [cs.LG].

[12] **Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman**. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." In: *Frontiers in Big Data* 2 (2019), p. 13. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00013.

[13] **Gordon Pennycook and David G. Rand**. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." In: *Cognition* 188 (2019). The Cognitive Science of Political Thought, pp. 39–50. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition

.2018.06.011. URL: https://www.sciencedirect.com/science/article/pii/S001002771830163X.

[14] **Jon Roozenbeek and Sander van der Linden**. "Fake news game confers psychological resistance against online misinformation." In: *Palgrave Communications* 5.1 (2019), p. 65. DOI: 10.1057/s41599-019-0279-9. URL: https://doi.org/10.1057/s41599-019-0279-9.

[15] **Tanik Saikh, Amit Anand, Asif Ekbal, and Pushpak Bhattacharyya**. "A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features." In: June 2019, pp. 345–358. ISBN: 978-3-030-23280-1. DOI: 10.1007/978-3-030-23281-8_30.

[16] **Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu**. *Combating Fake News: A Survey on Identification and Mitigation Techniques*. 2019. arXiv: 1901.06437 [cs.LG].

[17] **Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu**. "Unsupervised Fake News Detection on Social Media: A Generative Approach." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 5644–5651. DOI: 10.1609/aaai.v33i01.33015644.

[18] **Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung**. "Detecting Incongruity between News Headline and Body Text via a Deep Hierarchical Encoder." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 791–800.

[19] **Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal**. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2018. arXiv: 1806.00069 [cs.AI].

[20] **Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali**. "Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape." In: *CoRR* abs/1803.10124 (2018). arXiv: 1803.10124. URL: http://arxiv.org/abs/1803.10124.

[21] **Srijan Kumar and Neil Shah**. *False Information on Web and Social Media: A Survey*. 2018. arXiv: 1804.08559 [cs.SI].

[22] **Nicole M. Lee**. "Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom." In: *Communication Education* 67.4 (2018), pp. 460–466. DOI: 10.1080/03634523.2018.1503313. eprint: https://doi.org/10.1080/03634523.2018.1503313. URL: https://doi.org/10.1080/03634523.2018.1503313.

[23] **Yang Liu and Yi-fang Brook Wu**. "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks." In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018), pp. 354–361.

[24] **Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum**. "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning." In: *EMNLP*. 2018, pp. 22–32.

[25] **Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein**. "The Clickbait Challenge 2017: Towards a Regression Model for Clickbait Strength." In: *CoRR* abs/1812.10847 (Dec. 2018). URL: https://arxiv.org/abs/1812.10847.

[26] **Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein**. "A Stylometric Inquiry into Hyperpartisan and Fake News." In: *ACL*. Vol. 1. 2018, pp. 231–240.

[27] **Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu**. "Neural User Response Generator: Fake News Detection with Collective User Intelligence." In: IJCAI '18. 2018, pp. 3834–3840.

[28] **K. Shu, S. Wang, T. Le, D. Lee, and H. Liu**. "Deep Headline Generation for Clickbait Detection." In: *2018 IEEE International Conference on Data Mining (ICDM)*. 2018, pp. 467–476.

[29] **Kai Shu, Suhang Wang, and Huan Liu**. "Understanding User Profiles on Social Media for Fake News Detection." English (US). In: Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018. June 2018, pp. 430–435. DOI: 10.1109/MIPR.2018.00092.

[30] **James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal**. "FEVER: a Large-scale Dataset for Fact Extraction and VERification." In: New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074.

[31] **Svitlana Volkova and Jin Yea Jang**. "Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media." In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France, 2018, pp. 575–583. ISBN: 9781450356404. DOI: 10.1145/3184558.3188728.

[32] **Soroush Vosoughi, Deb Roy, and Sinan Aral**. "The spread of true and false news online." In: *Science* 359.6380 (2018), pp. 1146–1151. ISSN: 0036-8075. DOI: 10.1126/science.aap9559. eprint: https://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: https://science.sciencemag.org/content/359/6380/1146.

[33] **Hunt Allcott and Matthew Gentzkow**. "Social Media and Fake News in the 2016 Election." In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–36. DOI: 10.1257/jep.31.2.211. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[34] **Michelle A Amazeen**. "Journalistic interventions: The structural factors affecting the global emergence of fact-checking." In: *Journalism* 21 (2021/08/24 2017), pp. 95–111. DOI: 10.1177/1464884917730217. URL: https://doi.org/10.1177/1464884917730217.

[35] **John Cook, Stephan Lewandowsky, and Ullrich K. H. Ecker**. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence." In: *PLOS ONE* 12.5 (May 2017), pp. 1–21. DOI: 10.1371/journal.pone.0175799. URL: https://doi.org/10.1371/journal.pone.0175799.

[36] **Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach**. "Inoculating the Public against Misinformation about Climate Change." In: *Global Challenges* 1.2 (2017),

p. 1600008. DOI: https://doi.org/10.1002/gch2.201600008. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.201600008.

[37] **Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea**. "Automatic detection of fake news." In: *arXiv preprint arXiv:1708.07104* (2017).

[38] **Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum**. "Where the truth lies: Explaining the credibility of emerging claims on the web and social media." In: *WWW*. 2017, pp. 1003–1012.

[39] **Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf**. "BaitBuster : A Clickbait Identification Framework." In: (2017), pp. 8216–8217. arXiv: arXiv:1607.04606.

[40] **Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu**. "Fake News Detection on Social Media: A Data Mining Perspective." In: *SIGKDD Explor. Newsl.* 19.1 (Sept. 2017), pp. 22–36. ISSN: 1931-0145. DOI: 10.1145/3137597.3137600.

[41] **William Yang Wang**. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: https://www.aclweb.org/anthology/P17-2067.

[42] **Hai-Tao Zheng, Xin Yao, Yong Jiang, Shu-Tao Xia, and Xi Xiao**. "Boost Clickbait Detection Based on User Behavior Analysis." In: *Web and Big Data*. Ed. by Lei Chen, Christian S. Jensen, Cyrus Shahabi, Xiaochun Yang, and Xiang Lian. Cham: Springer International Publishing, 2017, pp. 73–80.

[43] **Chang Zhou, Yuqiong Liu, Xiaofei Liu, Zhongyi Liu, and Jun Gao**. "Scalable Graph Embedding for Asymmetric Proximity." In: AAAI'17. 2017.

[44] **Yiwei Zhou**. *Clickbait Detection in Tweets Using Self-attentive Network*. 2017. arXiv: 1710.05364 [cs.CL].

[45] **A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly**. "Stop Clickbait: Detecting and preventing clickbaits in online news media." In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016, pp. 9–16.

[46] **Aditya Grover and Jure Leskovec**. "Node2vec: Scalable Feature Learning for Networks." In: KDD '16. 2016, pp. 855–864.

[47] **Oberauer K Lewandowsky S**. "Motivated Rejection of Science." In: *Current Directions in Psychological Science* 25(4) (2016), pp. 217–222. DOI: 10.1177/0963721416654436.

[48] **Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha**. "Detecting Rumors from Microblogs with Recurrent Neural Networks." In: *IJCAI*. 2016, pp. 3818–3824.

[49] **Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko**. *Stance and Sentiment in Tweets*. 2016. arXiv: 1605.01655 [cs.CL].

[50] **Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit**. "A Decomposable Attention Model for Natural Language Inference." In: (2016). ISSN: 0001-0782. eprint: 1606.01933.

[51] **Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen**. "Clickbait Detection." In: *Advances in Information Retrieval*. Ed. by Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello. 2016, pp. 810–817. ISBN: 978-3-319-30671-1.

[52] **Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang**. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://www.aclweb.org/anthology/D16-1264.

[53] **Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy**. "Hierarchical Attention Networks for Document Classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174.

[54] **Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie**. "Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads." In: *PLOS ONE* 11 (Mar. 2016), pp. 1–29. DOI: 10.1371/journal.pone.0150989. URL: https://doi.org/10.1371/journal.pone.0150989.

[55] **Jonas Nygaard Blom and Kenneth Reinecke Hansen**. "Click bait: Forward-reference as lure in online news headlines." English. In: *Journal of Pragmatics* 76 (Jan. 2015), pp. 87–100. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2014.11.010.

[56] **Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning**. "A large annotated corpus for learning natural language inference." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://www.aclweb.org/anthology/D15-1075.

[57] **Yimin Chen, Niall J. Conroy, and Victoria L. Rubin**. "Misleading Online Content: Recognizing Clickbait as "False News"." In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. WMDD '15. Seattle, Washington, USA: Association for Computing Machinery, 2015, pp. 15–19. DOI: 10.1145/2823465.2823467.

[58] **Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong**. "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites." In: CIKM '15. Melbourne, Australia, 2015, pp. 1751–1754.

[59] **Brendan Nyhan and Jason Reifler**. "The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators." In: *American Journal of Political Science* 59.3 (2015), pp. 628–640. ISSN: 00925853, 15405907. URL: http://www.jstor.org/stable/24583087.

[60] **Duyu Tang, Bing Qin, and Ting Liu**. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1422–1432. DOI: 10.18653/v1/D15-1167.

[61] **Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei**. "LINE: Large-Scale Information Network Embedding." In: WWW '15. 2015.

[62] **Ke Wu, Song Yang, and Kenny Q. Zhu**. "False rumors detection on Sina Weibo by propagation structures." In: *2015 IEEE 31st International Conference on Data Engineering*. 2015, pp. 651–662. DOI: 10.1109/ICDE.2015.7113322.

[63] **Zhe Zhao, Paul Resnick, and Qiaozhu Mei**. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts." In: *WWW*. 2015, pp. 1395–1405.

[64] **Zhe Zhao, Paul Resnick, and Qiaozhu Mei**. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405. ISBN: 9781450334693. DOI: 10.1145/2736277.2741637.

[65] **Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio**. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680.

[66] **Yoon Kim**. "Convolutional neural networks for sentence classification." In: *arXiv preprint arXiv:1408.5882* (2014).

[67] **Jeffrey Pennington, Richard Socher, and Christopher D. Manning**. "Glove: Global vectors for word representation." In: *In EMNLP*. 2014.

[68] **Bryan Perozzi, Rami Al-Rfou, and Steven Skiena**. "DeepWalk: Online Learning of Social Representations." In: KDD '14. 2014, pp. 701–710.

[69] **Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu**. *Social Media Mining: An Introduction*. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139088510.

[70] **Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean**. "Distributed Representations of Words and Phrases and Their Compositionality." In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[71] **Sadia Afroz, Michael Brennan, and Rachel Greenstadt**. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online." In: *2012 IEEE Symposium on Security and Privacy*. 2012, pp. 461–475. DOI: 10.1109/SP.2012.34.

[72] **Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor**. "Cats Rule and Dogs Drool!: Classifying Stance in Online Debate." In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 1–9.

[73] **Carlos Castillo, Marcelo Mendoza, and Barbara Poblete**. "Information credibility on twitter." In: *WWW*. 2011.

[74] **José Luis Iribarren and Esteban Moro**. "Affinity Paths and information diffusion in social networks." In: *Social Networks* 33.2 (May 2011), pp. 134–142. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2010.11.003.

[75] **Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock**. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 309–319. URL: https://www.aclweb.org/anthology/P11-1032.

[76] **IDO DAGAN, BILL DOLAN, BERNARDO MAGNINI, and DAN ROTH**. "Recognizing textual entailment: Rational, evaluation and approaches." In: *Natural Language Engineering* 15.4 (2009), pp. i–xvii. DOI: 10.1017/S1351324909990209.

[77] **Danny Hayes**. "Echo Chamber: Rush Limbaugh and the Conservative Media Establishment by Kathleen Hall Jamieson and Joseph N. Cappella." In: *Political Science Quarterly* 124.3 (2009), pp. 560–562. DOI: https://doi.org/10.1002/j.1538-165X.2009.tb01921.x.

[78] **Rada Mihalcea and Carlo Strapparava**. "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language." In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACLShort '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 309–312.

[79] **Richard M. Perloff**. "The Dynamics of Persuasion: Communication and Attitudes in the 21st Century (2nd ed.). Routledge." In: 2007, p. 424. ISBN: 978-3-030-23280-1. DOI: https://doi.org/10.4324/9781410606884.

[80] **Xiaoxin Yin, Jiawei Han, and Philip S. Yu**. "Truth Discovery with Multiple Conflicting Information Providers on the Web." In: KDD '07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 1048–1052. ISBN: 9781595936097. DOI: 10.1145/1281192.1281309.

[81] **Gary Bond and Adrienne Lee**. "Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language." In: *Applied Cognitive Psychology* 19 (Apr. 2005), pp. 313–329. DOI: 10.1002/acp.1087.

[82] **Alex Graves, Santiago Fernández, and Jürgen Schmidhuber**. "Bidirectional LSTM networks for improved phoneme classification and recognition." In: *ANN*. Springer. 2005, pp. 799–804.

[83] **Theresa Wilson and Janyce Wiebe**. "Annotating attributions and private states." In: (July 2005), pp. 53–60. DOI: 10.3115/1608829.1608837.

[84] **Chin-Yew Lin**. "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[85] **Mark G. Frank and Thomas Hugh Feeley**. "To Catch a Liar: Challenges for Research in Lie Detection Training." In: *Journal of Applied Communication Research* 31.1 (2003), pp. 58–75. DOI: 10.1080/00909880305377. eprint: https://doi.org/10.1080/00909880305377.

[86] **Sepp Hochreiter**. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6 (Apr. 1998), pp. 107–116. DOI: 10.1142/S0218488598000094.

[87] **Sepp Hochreiter and Jürgen Schmidhuber**. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

# Paper I:
# SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection

# SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection

**Rahul Mishra**[1], **Vinay Setty**[1]

[1] University of Stavanger
Stavanger, Norway
[2] Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway

Published in 2019, The 9th ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19).

**Abstract:**
Recently false claims and misinformation have become rampant in the web, affecting election outcomes, societies and economies. Consequently, fact checking websites such as snopes.com and politifact.com are becoming popular. However, these websites require expert analysis which is slow and not scalable. Many recent works try to solve these challenges using machine learning models trained on a variety of features and a rich lexicon or more recently, deep neural networks to avoid feature engineering.
In this paper, we propose hierarchical deep attention networks to learn embeddings for various latent aspects of news. Contrary to existing solutions which only apply word-level self-attention, our model jointly learns the latent aspect embeddings for classifying false claims by applying hierarchical attention. Using several manually annotated high quality datasets such as Politifact, Snopes and Fever we show that these learned aspect embeddings are strong predictors of false claims. We show that latent aspect embeddings learned from attention mechanisms improve the accuracy of false claim detection by up to 13.5% in terms of Macro F1 compared to a state-of-the-art attention mechanism guided by claim-text (DeClarE). We also extract and visualize the evidence from the external articles which supports or disproves the claims.

# 1 Introduction

The unprecedented growth of the web, online news and social media has led to a paradigm shift in the way people consume information. As a consequence, spread of misinformation or fake news in online media has become faster and wider than ever before. To counter this, several fact checking websites such as snopes.com, politifact.com and fullfact.org are becoming increasingly popular. These websites have dedicated experts manually classifying the credibility of news articles and claims which is slow and tedious.

To address these limitations, several automated machine learning models are proposed in the literature. Early works in this area focused on the tedious task of curating a rich lexicon and other credibility features manually to capture the language of deception [11, 5, 12]. More recent approaches avoid feature engineering by designing deep neural network models which are able to learn non-trivial patterns from the raw text of the claims or facts [13]. However, verifying correctness of claims purely based on the claim text has limited effectiveness due to lack of context information.

To overcome the above problem, recent works incorporate ***external evidence*** retrieved from news media and social media which potentially either supports or refutes the claim [4, 9]. These works propose a word-level attention mechanism guided by the claim text to focus on parts of the external evidence for this purpose. However, it has been shown that word-level attention alone fails to capture the complex structure of the documents [18]. **Hierarchical attention** mechanism which applies attention at sentence level in addition is shown to be more effective for document classification. For example, in Figure 1, we can notice that using word-level attention (in red font) alone makes it hard to determine if the evidence supports or refutes the claim. However, the sentence level attention (highlighted text) is able to capture the context better.

While attention guided by the claim text is effective to some extent for detecting fake news, it has been shown that it is not sufficient [4]. In order to effectively use external evidence for fake news detection, determining its credibility in the context of the given claim and its author (source) is also essential. Popat et. al in [4], propose the use of static representation

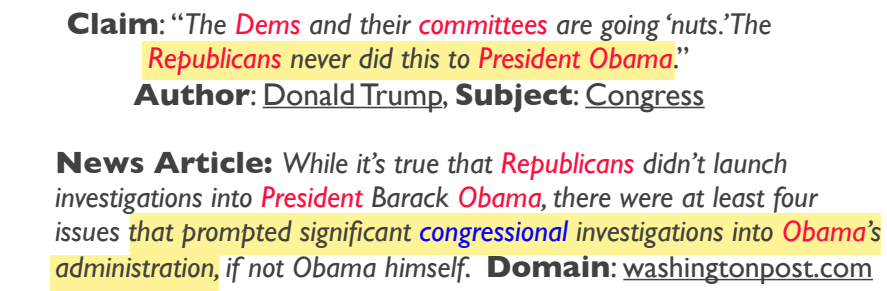



Figure 1: Example of a false claim, word and sentence-level attention using latent aspects (Subject, Author and Domain)

(one-hot encoding) of source information (domains) together with attention weights for this purpose. However, we postulate that understanding the context and credibility of the evidence requires learning indicative and salient vocabulary, writing style and sentence structures specific to the latent aspects of news. For example, in Figure 1, relying only on the *claim-text attention* does not successfully classify that it is a false claim. Given that the claim is from "Donald Trump" (**Author**), related to the **Subject** "Congress" can guide the attention to a new word such as "congressional" which is missing in the claim text. In addition, the professional writing style of journalists from the **Domain** "washingtonpost.com" further refutes the claim. We hypothesize that the attention due to latent aspects is able to capture the necessary patterns to check if the external evidence supports or refutes the claim. This task is commonly known as **entailment**.

To address these limitations, in this paper, we propose a novel model coined SADHAN[1] to jointly learn embeddings for different latent aspects of news using the hierarchical attention mechanism. Intuitively, the attention mechanism learns a unique global representation (embedding) for each of the latent aspects. These embeddings capture the necessary textual patterns needed to distinguish a claim for being true or false. For example, an embedding learned for the author "Donald Trump" captures the patterns from discussions about false and true claims made by him. Similarly, embeddings representing each of the latent aspects capture the

[1]Subject, Author, Domain Based Hierarchical Attention Network

necessary patterns from the representative relevant articles to distinguish the veracity of the claims. We illustrate that these embeddings are indeed able to distinguish false and true news in Section 6 by visualizing these embeddings in two dimensions using t-SNE. Note that the latent aspect embeddings are not limited to the subject, author and domain aspects but they are general purpose and can be used for any latent aspects which are relevant for the task.

One of the critical tasks performed by the fact checking websites is to provide evidence for the veracity of the claim. This is a highly cognitive task and usually done manually by experts. Therefore, it is not sufficient to just automate fake news detection but also to automatically extract the supporting evidence. In previous works, word-level attention weights are used to extract the evidence and visualize the words in textual snippets [4]. However, just using word level attention weights to visualize evidence is not very user friendly. In this paper, we propose an algorithm to fuse the word level and sentence level attention weights guided by various latent aspect embeddings to extract evidence snippets which are easier for humans to understand.

In summary, our contributions are:

(1) Hierarchical attention to learn claim and document structure

(2) Jointly learning latent aspects of news using hierarchical attention mechanism

(3) Extensive experiments using three high quality datasets

(4) Visualization and analysis of latent aspect embeddings

(5) Evidence extraction and visualization of attention mechanism for interpretability

Our experiments using data crawled from Snopes and Politifact, show that latent aspect embeddings jointly learned using SADHAN, are very effective in detecting false claims. Specifically, we gain up to 12% improvement in Macro F1 for Politifact and 13.5% for Snopes compared to the state-of-the-art solution based on claim text attention and source embeddings [4]. In addition, we illustrate that the latent aspect embeddings learned by our model are effective for detecting false claims on their own

by visualizing them.

## 2 Related Work

Some of the first methods for detecting fake news have been using linguistic cues [11, 5] and source-based credibility features [12]. However, identifying the specific linguistic cues that are decisive for fake news is not yet fully understood.

Deep learning methods to avoid feature engineering have also been proposed [4, 13, 3, 1]. In [4], the authors concatenate the claim text and content of the article and apply word-level self-attention to detect false claims. We improve on this architecture to include sentence-level attention as well as attention guided by the latent aspect embeddings. In [4], the authors also use word-level attention to extract evidence snippets, we instead propose an algorithm to select top-K sentences based on the attention weights both at word and sentence level.

There are also efforts to address some sub-problems of detecting false claims such as entailment [9, 7]. Another related task is stance detection[2]. While these tasks are not the same as detecting fake news, it could be used for checking the veracity of claims. SADHAN implicitly also depends on entailment among other patterns to detect fake news. To support this, we also use our model to evaluate the Fever dataset published by [7].

Several recent works have shown that using context from social network users and interactions have improved fake news detection [2, 19, 16, 8, 15, 6]. However, these approaches are only suitable when there is sufficient information from social networks associated with the news available. We could integrate SADHAN into these models to achieve further improvements.

The neural architecture of SADHAN is inspired by the hierarchical architecture in [18] originally proposed for document classification. While speaker-based attention has been used before [10], using it for hierarchical attention and multiple latent aspects has never been tried before for fake news detection.

[2] http://www.fakenewschallenge.org

In summary, we are the first to propose a hierarchical attention network which jointly learns latent aspect embeddings to detect fake news.

# 3 Problem Definition and Proposed Model

## 3.1 Problem Definition

Given a claim $c \in C$ in textual form, along with its latent aspects such as subject, author, domain and a set of candidate relevant documents $D = \{d_1, ...., d_m\}$ as evidence from different domains, the goal is to classify the claim as either "True" or "False".

## 3.2 SADHAN Model

Now we explain the SADHAN model in detail. The overall architecture is depicted in Figure 2 upper section. Given a training dataset of claims with their ground-truth labels, our goal is to learn a model based on the evidence from the relevant web documents $D$. To address the two challenges mentioned in Section 1, (1) we use a **hierarchical Bi-LSTM** model to capture the word-level and sentence-level structure of the documents, (2) An attention mechanism, which uses both claim text and latent aspect attribute vector to compute the attention, is then used to learn the embedding weights of the latent aspects. The intuition behind this design is that each of the latent aspect models jointly guide the attention to vocabulary and the sentences relevant for classifying claims. This architecture as we show in the experiments learns an effective model to identify complex patterns of false claims. For this purpose, SADHAN has different parallel models, one for each of the latent aspects. The detailed architecture of these models is shown in Figure 2 (zoomed in lower section). Specifically, we consider Subject, Author and Domain aspects in this paper but it is generalizable to any additional aspects of the claims and documents. At a high level, each claim-document pair $\{c, d\}$ is passed as the input to each of the three models, along with respective latent aspects. The outputs of these models are concatenated and passed to a fully connected softmax layer for prediction. Losses of all three models are aggregated using a noisy-or function. Finally, since our models operate on claim-document

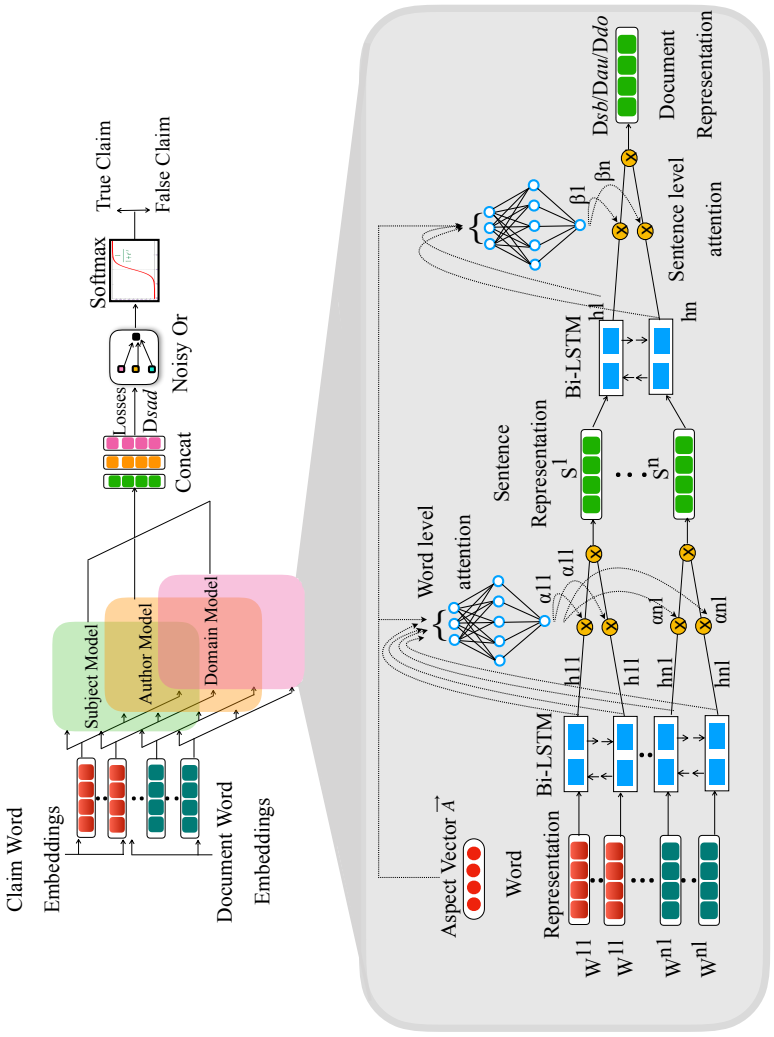



Figure 2: high-level architecture of SADHAN (upper part) and detailed hierarchical attention architecture of latent aspect models (lower part)

pairs, the classification of the claim $c_i$ is done by the majority voting of outcomes corresponding to each of the $\{c, d\}$ pairs. We now explain the architecture of SADHAN in detail.

**Embedding Layer:** We use pretrained GloVe embeddings to get representations for each claim and document pair. We also create trainable embeddings for subject, author and domain attributes of 100 dimensions each in size and initialize with uniformly random weights to get the representation of latent attributes in vector space. We learn weights for these embeddings jointly in the model using corresponding hierarchical subject, author and domain attentions from their respective models as shown in Figure 2 (lower part). We concatenate each embedded claim $c_i$ with the corresponding embedded document $d_j$, which is denoted as $\{c, d\}$. Each $\{c, d\}$ pair consists of $n$ sentences of length $l$, which is depicted as word sequence $w^{11}$ to $w^{nl}$ in Figure 2 (lower part).

## 3.3 Latent Aspect Attention

Different authors, while making claims on different subjects, tend to have different styles of speech and selection of words. Similarly, writers and journalists from different domains may have unique style and vocabulary while writing about claims from a specific author and a specific subject. It is an extremely difficult task to curate the salient vocabulary and sentence structures for these complex combinations. Therefore, we automate this task using an attention mechanism which in turn helps in capturing entailment and sentiments necessary to classify the claim. For example, in tweets by Donald Trump words like "great", "democrats" and "obama" are normally mentioned in specific context and sentiments, which our attention mechanism is able to capture.

Each claim and document pair $\{c, d\}$ is associated with a subject vector $\vec{A}_s$, author vector $\vec{A}_a$ and domain vector $\vec{A}_d$. These aspect vectors are used in addition to claim text to learn attention weights applied to hidden states at both word level and sentence level. The concatenated word embeddings of claim and document pair $\{c, d\}$ are passed on to a Bi-directional LSTM [21] unit which we use as word encoder, output from these Bi-LSTM units are concatenations of forward and backward hidden states for each word.

$h_{ij}$ is the hidden state for the $i^{th}$ word of the $j^{th}$ sentence. We compute values of attention weights $\alpha^{11}$ to $\alpha^{nl}$ by using single layer neural net with $\tanh$ activation, which uses encoded hidden states of claim and doc pair and aspect attribute vector $\vec{A}$ as input. We then multiply these attention weights $\alpha^{11}$ to $\alpha^{nl}$ with corresponding hidden states to select significant words, which are used to form sentence representations as $s^1$ to $s^n$. These sentence representations are then processed by another Bi-LSTM layer, which outputs hidden states $h_1$ to $h_n$ for each sentence, as shown in the Figure 2(lower part). We compute values of attention weights $\beta^1$ to $\beta^n$ by using another single layer neural net with $\tanh$ activation, which uses hidden states of sentences and aspect attribute vector $\vec{A}$ as input. We then multiply these attention weights $\beta^1$ to $\beta^n$ with corresponding hidden states of sentences to select significant sentences, which are used to form document representations as $D_{sb}/D_{au}/D_{do}$ in case of subject, author or domain models correspondingly.

**Subject Model:** The words which are significant for a specific subject, can be used in various ways by different authors in claims and by different columnists or journalists in articles related to claims, therefore subject attention at the word level tries to learn and attend these words and at the sentence level tries to capture significant sentence formations used for the specific subject.

**Author Model:** Similar to the subject model, we use author guided aspect attention at word level to select author related words used in articles and sentence representations are learned by aggregating these words. We apply author guided aspect attention at the sentence level to select author specific sentence formations or popular phrases which are frequently used for a specific author and we get document representation $D_{au}$ by aggregating these selected sentences.

**Domain Model:** Different domains in the web search results may have a unique way of writing articles like selection of words and sentence formations. In similar fashion to subject and author aspect attention, to attend different domains differently and to learn latent patterns, we apply domain guided aspect attention at the word and sentence level and get

document representation $D_{do}$.

More formally, in all three models, sentence representation $S^i$ after word sequence encoding by the Bi-LSTM is the weighted sum of the hidden states of words multiplied by attention weights. Similarly, document representation $D$ is the weighted sum of hidden states of sentences multiplied by attention weights. These are defined as:

$$S^i = \sum_{j=1}^{l_i} \alpha_{ij} h_{ij}$$

$$D = \sum_{i=1}^{n} \beta_i h_i$$

Where $h_{ij}$ is the hidden state for the $j^{th}$ word and $i^{th}$ sentence. $\alpha_{ij}$ is the attention weight. $h_i$ is the hidden state for $i^{th}$ sentence and $\beta_i$ is the attention weight. $\alpha_{ij}$ and $\beta_i$ can be defined as:

$$\alpha_{ij} = \frac{\exp(e(h_{ij}, \vec{A}))}{\sum_{k=1}^{l_i} \exp(e(h_{ik}^s, \vec{A}))}$$

$$\beta_i = \frac{\exp(e(h_i, \vec{A}))}{\sum_{k=1}^{n} \exp(e(h_k, \vec{A}))}$$

Where $e$ is a $tanh$ based scoring function which decides weights for significant words at the word level attention and for significant sentences at sentence level attention. $\vec{A}$ is the latent aspect vector, which is equal to subject vector $\vec{A}_s$ in subject model, author vector $\vec{A}_a$ in author model and domain vector $\vec{A}_d$ in case of domain model. $e(h_{ij}, \vec{A})$ and $e(h_i, \vec{A})$ can be defined as:

$$e(h_{ij}, \vec{A}) = (v_w)^T \tanh(W_{wh} h_{ij} + W_{wA} \vec{A} + b_w)$$

$$e(h_i, \vec{A}) = (v_s)^T \tanh(W_{sh} h_{ij} + W_{sA} \vec{A} + b_s)$$

Where $v_w$ is weight vector at the word level and $v_s$ is weight vector at the sentence level. $W_{wh}$ and $W_{wA}$ are the weight matrices for hidden state and aspect vector and $b_w$ is bias at the word level respectively. $W_{sh}$ and $W_{sA}$ are the weight matrices for hidden state and aspect vector and $b_s$ is bias at the sentence level respectively.

## 3.4 Fusion of Models

Representations for each document $D$ are learned from all three models as $D_{sb}$ from subject model, $D_{au}$ from author model and $D_{do}$ from domain model. We concatenate these three representations for the same document and form an overall representation. $D_{sad} = D_{sb} \oplus D_{au} \oplus D_{do}$ We apply a non-linear transformation on overall document representation $D_{sad}$ using $\tanh$ dense layer to transform it to binary target space. $D_{bin} = \tanh(W_{bin}D_{sad} + b_{bin})$ where $W_{bin}$ and $b_{bin}$ are the weight matrix and bias for dense layer. We apply a softmax layer to obtain the predictions for each class $P_{bin}$ as $P_{bin} = softmax(D_{bin})$. Finally, we combine the losses of all three models with noisy-or gate as below:

$$Loss = 1 - ((1 - loss_o)) * (1 - loss_s) * (1 - loss_a) * (1 - loss_d))$$

where $loss_o$, $loss_s$,$loss_a$ and $loss_d$ are the losses for overall merged model, subject model, author model and domain model respectively.

## 3.5 Prediction Per Claim

The prediction outcomes for a claim $c$ paired with each corresponding documents $\{d_1, ...., d_m\}$ are then aggregated by majority voting to assign a class to the claim.

$$\hat{y} = mode\{y_1, y_2, ...y_m\}$$

Where $\hat{y}$ is the final predicted label for claim $c$ and $y_1, y_2, ...y_m$ are the predictions for pairs of claim $c$ and corresponding $m$ documents.

## 3.6 Evidence Extraction

In this section, we propose a technique to extract evidence snippets supporting or refuting the claim from documents using attention weights at both the word and sentence level from all three models. The pseudocode is shown in tab:algo. In line 5, for each word in each sentence of document $d$, we compute the average of attention weights given by all three models and this gives us overall attention weight for that word. In line 7, we compute the average of overall attention weights for all words in

**Algorithm 1** Evidence Extraction Algorithm

**Input:** Claim $c \in C$; Document $d \in D$; $W_{ws}, W_{wa}, W_{wd}$ are the word level and $W_{ss}, W_{sa}, W_{sd}$ are the sentence level attention weight matrices for subject, author and domain model respectively; $K$ is number of sentences in evidence snippet

**Output:** $E$, an evidence snippet for claim $c$

1 $S = []$ `// Initialize an empty list`
2 **for** *each sentence* $s_i$ *in* $d$ **do**
3 $\quad W = []$ `// Initialize an empty list`
4 $\quad$ **for** *each word* $w_{ij}$ *in* $s_i$ **do**
5 $\quad\quad W.append((W_{ws}[i,j] + W_{wa}[i,j] + W_{wd}[i,j])/3)$
6 $\quad$ **end**
7 $\quad W_{avg} \leftarrow sum(W)/len(W)$ $S[i] \leftarrow W_{avg} + (W_{ss}[i] + W_{sa}[i] + W_{sd}[i])/3$
8 **end**
9 $indexes \leftarrow argsort(S)[-K:]$ `// Get indices of top K elements from` $S$
10 $E \leftarrow d[indexes]$ `// Get sentences corresponding to indices from` $d$
11 **return** $E$

a sentence and add this value to the average of sentence level attention weights for the same sentence from all three models and store this value to list $S$. We get indices of top K values in $S$ using $argsort$ (line 9) and get the corresponding sentence indices from document $d$ (line 10).

# 4 Experimental Setup

## 4.1 Datasets

We use three datasets–Politifact and Snopes released by Popat et al [4] and Fever dataset released by Thorne et al [7].

**Politifact Dataset** Politifact has 3568 claims and 29556 documents associated with 3028 domains retrieved from the web search using Bing search API. We discard articles related to fact-checking domains. For each claim, Politifact has one of these six ratings: 'true', 'mostly true', 'half true', 'mostly false', 'false' and 'pants-on-fire'. Similarly to DeClarE we

combine 'true', 'mostly true' and 'half true' ratings to 'true' label and rest of them to 'false' label. There are $669$ unique authors and $1400$ topics in total.

**Snopes Dataset** Snopes has around $4341$ claims and $29242$ documents associated with $3267$ domains retrieved from the web using Bing search API. Similar to Politifact we discard all the documents which are from fact checking websites such as Snopes, Politifact, Factcheck and Emergent etc. For each claim, it has a credibility label as 'True' or 'False'.

**Fever Dataset** While Fever dataset is not dedicated for fake news detection in itself, it is widely used for the entailment task, which can be viewed as a subtask of fake news detection. We use the fever dataset to illustrate that our model is also effective for the entailment task. This is to validate our hypothesis that our model performs well because it is also able to perform entailment task effectively. Fever dataset has $145449$ claim-evidence pairs in trainset, $9999$ claim-evidence pairs in development set and $9999$ claim-evidence pairs in test set (for more details see [7]). In addition to what is already present in Fever dataset, we use Latent Dirichlet Allocation (LDA) to get the dominant topic for each claim in train, validation and test dataset as Fever dataset doesn't have any aspect attributes. We use the elbow method with topic coherence score to tune the number of topics $K$, as a result we use $K = 273$.

### 4.1.1 Data Imbalance

Since Snopes and Politifact datasets have class imbalance, we balance them by setting the `class_weight` parameter to "balanced" in scikit–learn `compute_class_weight` API[3]. On the other hand Fever dataset is already balanced.

## 4.2 Baselines

We compare our model using several baselines both simple baselines and state-of-the-art techniques:

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html`

(1) Simple Convolutional Neural Network (CNN) model which was proposed for sentence classification [20]

(2) Hierarchical LSTM Network (Hi-LSTM) for document classification (without attention) [18]

(3) Self-attention based Hierarchical Attention Network BiLSTM (HAN) [18]

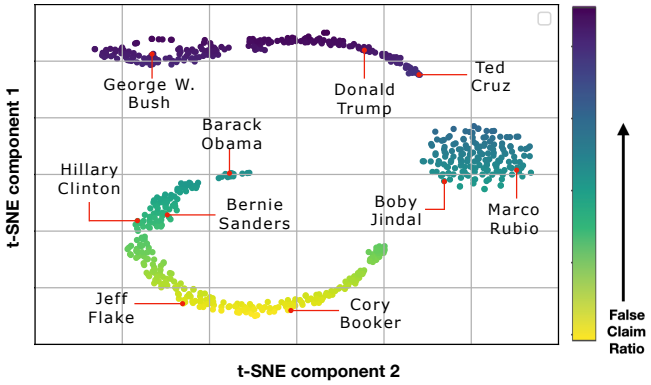
(4) DeClarE which applies claim-text based attention and source based embeddings [4]

To perform ablation testing for our SADHAN model, we incrementally introduce various latent aspect embeddings over the HAN architecture. We represent our models as SHAN, AHAN, and DHAN for each of the latent aspects Subject, Author and Domain respectively. Finally, SADHAN is our full model with all three aspects. Each of these models perform classification at the document level. DeClarE on the other hand performs classification on a per claim basis. Therefore in order to compare the performance of our model to DeClarE we also evaluate an aggregated version of our model represented as SADHAN-agg, which uses mean score from predictions of individual articles to assign a class to the claim.

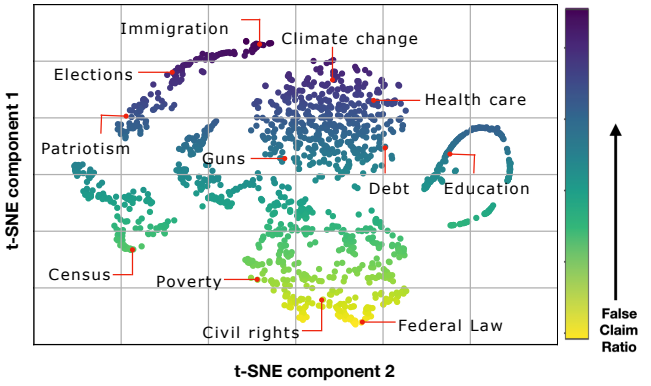
## 4.3 SADHAN Implementation

We implement SADHAN using TensorFlow framework. We use 10 fold cross validation for all the models. We compute per-class accuracy, Macro F1 score and AUC as performance metrics for evaluation. We use pre-trained GloVe embeddings of $100$ dimensions, trained on $6$ billion words. We extract relevant snippets of text from the web documents using cosine similarity to include only highly relevant parts of the web documents. We try different sentence length sizes but we see no noticeable difference in performance. We tune the parameters [4] using a validation set, as a result we use softmax cross entropy with logits as the cost function, learning rate of $0.001$ and size of hidden states and cell states of Bi-LSTM units are kept as $200$. For drop out regularization we used keep-prob $= 0.3$. We chose these hyperparameters via grid search.

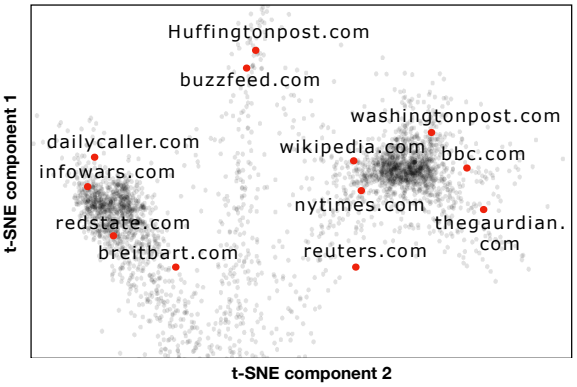
[4] `https://github.com/rahulOmishra/SADHAN`




(a) Author Embedding Visualization




(b) Subject Embedding Visualization




(c) Domain Embedding Visualization

Figure 3: Visualization of Latent Embeddings (The darker the color higher the false claim ratio).

# 5 Experimental Results

## 5.1 Results for Politifact Dataset

In Table 5.2 for Politifact dataset, in case of CNN, we get $59.39\%$ Macro $F1$ accuracy and $58.56\%$ as $AUC$. Hi-LSTM performs slightly better than CNN with $60.11\%$ Macro $F1$ accuracy and $60.66\%$ as $AUC$, though we get better false class accuracy with the Hi-LSTM. The reason for this improvement is that Hi-LSTM captures the inherent hierarchical structure of the documents. On the other hand HAN performs significantly better than Hi-LSTM with $64.80\%$ Macro $F1$ accuracy and $64.54\%$ as $AUC$ and provides gain of $6.6\%$ in Macro $F1$ over Hi-LSTM. The reason for this is because the documents retrieved from the web are fairly large even after extracting only relevant snippets using cosine similarity technique. It is hard for LSTM networks to memorize such long sequences. Moreover, LSTM with Attention mechanism only remembers attended words at word level and only attended sentences at sentence level.

As Politifact dataset has all aspect attributes such as subject, author and domain, we apply all individual models. Each of the SHAN, AHAN and DHAN models outperform HAN in Macro $F1$ with Macro $F1$ as $65.36\%$, $66.83\%$ and $65.05\%$ respectively. AHAN performs slightly better than the other two. This is due to the fact that the subject aspects in Politifact are generic. For each domain in domain attribute, we have high variance because each domain might have articles written by many different writers having different writing styles. The full SADHAN model outperforms all the other models with significant gain of $7.5\%$ in Macro $F1$. This gain can be attributed to fusion of three models, which considers all aspects of the claim and document pair for classification.

## 5.2 Results for Snopes Dataset

For Snopes, we can see in Table 5.2 that Hi-LSTM with $74.33\%$ Macro $F1$ accuracy and $79.20\%$ as $AUC$ outperforms CNN with $72.63\%$ Macro $F1$ accuracy and $76.45\%$ as $AUC$ by $2.7\%$ in Macro $F1$ and similar to Politifact results, this gain is also attributed to better representation learned in the form of the hierarchical structure of the documents by Hi-LSTM. HAN with $77.80\%$ Macro $F1$ accuracy and $80.33\%$ as $AUC$ gives further

Table 5.1: Comparison of proposed model with various state of the art baseline models for False claim detection on Snopes and PolitiFact datasets

| Data | Model | True Acc. | False Acc. | Macro F1 | AUC |
|---|---|---|---|---|---|
| PolitiFact | CNN | 55.92 | 57.33 | 59.39 | 58.56 |
| | Hi-LSTM | 55.85 | 65.86 | 60.11 | 60.66 |
| | HAN | 60.32 | 68.20 | 64.80 | 64.54 |
| | SHAN | 62.29 | 68.43 | 65.36 | 65.23 |
| | AHAN | 63.25 | 70.42 | 66.83 | 68.66 |
| | DHAN | 60.34 | 69.76 | 65.05 | 65.03 |
| | SADHAN | **69.79** | **75.45** | **71.34** | **72.37** |
| Snopes | CNN | 72.05 | 74.29 | 72.63 | 76.45 |
| | Hi-LSTM | 74.21 | 74.16 | 74.33 | 79.20 |
| | HAN | 76.76 | 79.65 | 77.80 | 80.33 |
| | DHAN | **77.06** | **81.63** | **78.73** | **82.03** |

Table 5.2: Comparison of proposed model with DeClarE models for False claim detection on Snopes and PolitiFact datasets. SADHAN-agg is statistically significant ($p - value = 1.05e^{-4}, 2.45e^{-2}$ for Politifact and Snopes respectively using pairwise student's t-test)

| Data | Model | True Acc. | False Acc. | Macro F1 | AUC |
|---|---|---|---|---|---|
| PolitiFact | DeClarE (full) | 68.18 | 66.01 | 67.10 | 72.93 |
| | SADHAN-agg | **68.37** | **78.23** | **75.69** | **77.43** |
| Snopes | DeClarE (full) | 60.16 | 80.78 | 70.47 | 80.80 |
| | DHAN-agg | **79.47** | **84.26** | **80.09** | **85.65** |

gain of $4\%$ on top of Hi-LSTM, due to hierarchical attention at word and sentence level. Since Snopes dataset has only domain attribute, we only use (DHAN) with $78.73\%$ Macro $F1$ accuracy and $82.03\%$ as $AUC$, which outperforms all the baseline methods and gives gain of $1.2\%$ over HAN.

## 5.3 Evaluation of claim-level classification

Since DeClarE classifies claims rather than individual documents, we compare aggregated model SADHAN-agg with DeClarE (full) model which applies only claim-text based attention in Table 5.2.

For Politifact data, SADHAN-agg outperforms DeClarE (full) model by 12% in micro $F1$. We attribute these gains to the latent aspect level attention which is able to capture the context better. While only claim-text based attention learns to attend the words having connotation with claim at word level only.

For Snopes dataset, DHAN-agg with $80.09\%$ Macro $F1$ accuracy and $85.65\%$ as $AUC$ outperforms DeClarE (full) model with $70.47\%$ Macro $F1$ accuracy and $80.80\%$ as $AUC$ by 13.5% in micro $F1$. We attribute these gains to the usage of domain aspect attribute in addition to claim-text for attention computation.

## 5.4 Results for Fever Dataset

We used Fever dataset to investigate the effectiveness of our model for the textual entailment task. Since Fever data doesn't have any of the three subject, author or domain attributes, we use Latent Dirichlet Allocation (LDA) to get the dominant topic for each claim therefore we apply SHAN model for textual entailment. We get $79.20\%$ accuracy ($p - value = 3.62e^{-4}$ in pairwise student's t-test) with the testset and $83.09\%$ accuracy with devset provided with Fever dataset, which outperforms multi-layer perceptron (MLP) with $73.81\%$ accuracy (Riedel et al. 2017)[14] method used by authors of Fever dataset paper [7], which uses single hidden layer with TF-IDF vector based cosine similarity between the claim and evidence. On the other hand SHAN model could not outperform the decomposable attention model in [17] with $88.0\%$ accuracy. We hypothesize that this is because the derived dominant topics learned for claims using LDA topic model may not be a true representation of original topics of claims. We could improve the performance by using more concrete set of topics such as categories from Wikipedia.

**Claim:** *U.S. is 'most highly taxed nation in the world'*

| | Sentence Attention | Word Attention |
|---|---|---|
| **Author:** *Donald Trump* | | **1.** This was well below the UK which raised 36.3 per cent.<br>**2.** It was below Germany, which raised 45 per cent.<br>**3.** In fact of the 191 countries covered by the IMF data set the US was only the 71st most taxed country in the world by this metric.<br>**4.** Once various allowances and deductions are factored in, the effective corporate US rate on firms comes down to 18.6 per cent according to the US Congressional Budget Office. |
| **Domain:** *independent.co.uk* | | **1.** What about indirect taxes, like sales taxes or VAT etc? Again, Trump claim is grossly inaccurate by this this measure.<br>**2.** The US was actually the bottom of the list of 35 analysed countries.<br>**3.** this was lower than the rates seen in Japan and the UK.<br>**4.** And as for the idea of the American population as a whole being the most highly-taxed nation on earth that is simply a falsehood. |
| **Subject:** *Taxes* | | **1.** But was he talking about income tax rates? If he was, this was also a false claim<br>**2.** The US share of national income raised from corporate taxes is today considerably lower than in the 1960s and 70s when the US government raised upwards of 2.5 per cent of national income this way.<br>**3.** it is misleading to suggest that US firms are currently over-taxed in a practical sense.<br>**4.** And as for the idea of the American population as a whole being the most highly-taxed nation on earth that is simply a falsehood. |

**Output of Evidence Extraction**

This was well below the UK which raised 36.3 per cent. It was below Germany, which raised 45 per cent. this was lower than the rates seen in Japan and the UK. And as for the idea of the American population as a whole being the most highly-taxed nation on earth that is simply a falsehood. But was he talking about income tax rates? If he was, this was also a false claim.

**(a)**

| | |
|---|---|
| **DeClarE:** *Word attention* | **1.** total national income, or GDP, raised by the US government in tax.<br>**2.** In fact of the 191 countries covered by the IMF data set the US was only the 71st most taxed country in the world by this metric.<br>**3.** Again, thats well below the 25.8 per cent wedge in the UK, the 34 per cent wedge in Germany and the 40 per cent wedge in France.<br>**4.** And as for the idea of the American population as a whole being the most highly-taxed nation on earth that is simply a falsehood. |

**(b)**

Figure 4: Example 1: Comparison of SADHAN evidence extraction with DeClarE for the claim "U.S. is 'most highly taxed nation in the world'"

# 6 Discussion

In this section we analyze the effectiveness of latent aspect embeddings learned by our model and illustrate the interpretability of our model with the help of evidence extraction and attention visualization. We compare snippets extracted by our model to the attention visualization of DeClarE using anecdotal examples.

**Author Embeddings:** We use t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize author embeddings in lower dimensional space. We plot only two dimensions from t-SNE with tuned parameters ($perplexity = 10$, $learningrate = 0.1$ and $iterations = 2000$ ). We show the fraction of false claims associated with each author using a color gradient (cf. Figure 3). As we can see in the plot that the authors having a higher number of false claims are clearly separated from authors having a lower number of false claims. Interestingly we also notice the formation of a third cluster, which is related to the authors, who have almost equal number of false claims and true claims. This is also very interesting to see that people of similar ideology like 'Obama', 'Hillary' and 'Sanders' are closer in embedding space. This is evident by the visualization that the author based attention can distinguish very effectively between the authors with less connotation of false claims and the authors with high connotation of false claims, which in-turn helps in deciding the credibility of claims.

**Subject Embeddings:** Similarly, we plot two dimensions from t-SNE with tuned parameters ($perplexity = 20$, $learningrate = 1.0$ and $iterations = 3000$) to visualize the subject embeddings (cf. Figure 3(b)). We can observe in the plot that the subjects with low and high false claim ratios are separated clearly into clusters. Due to the coarser granularity of the subjects, the separation is not as pronounced as author embeddings. It is however, quiet insightful to see that the topics like 'Climate change' and 'Health care' have very high percentage of false claims and are closer in the two-dimensional space. While 'Federal law' which has very low associated false claims is far away from them.

**Domain Embeddings:** For domain embeddings, we use t-SNE with tuned parameters ($perplexity = 20$, $learningrate = 0.1$ and $iterations = 2000$ ) to plot two dimensions (Figure $3(c)$). Notice that the domain embeddings clearly separate trustworthy domains like 'washingtonpost.com', 'nytimes.com' etc. from non-trustworthy domains like 'inforwars.com' and 'dailycaller.com', making the learned domain embeddings good detectors of fake news.

## 6.1 Attention Visualization

In this section, we visualize the attention weights for two anecdotal examples (claim and document pairs), both at the word and sentence level for all three models and compare with state-of-the-art DeClarE model in Figure 4 and 5. The depth of the colors in rectangle boxes next to each sentence, represents the distribution of attention weights at the sentence level. Similarly depth of the color of highlights of the words represents the distribution of attention weights at the word level. For all the three models only top $4$ sentences in Figure 4 and top 2 sentences in Figure 5 based on both word and sentence level attention weights are shown. As in each of the three models we use both claim and document text on top of aspect attributes to compute attention therefore we get some common trends in both word level and sentence level attention for all three models. Due to usage of different aspect attributes namely subject, author and domain in different models for attention computation, we get very interesting and relevant words and sentences selected in all three, which is not possible otherwise.

As we can see in Figure 4(a), for a claim related to Donald Trump that "U.S. is the most highly taxed nation in the world'", we apply our model to detect if it's true or false. We use a document extracted from the web for which domain is "independent.co.in", author is "Donald Trump" and subject is "Taxes". In author model, we can observe that in Figure 4(a) first row, author based attention is able to capture words like "below Germany", "below the UK" and "Congressional Budget" other than claim oriented words like 'US' and 'Taxed' etc, as these words are highly correlated with the author "Donald Trump" as 'Germany', 'UK' and 'Congressional' are some of the frequent words used by 'Donald Trump' or can be found in

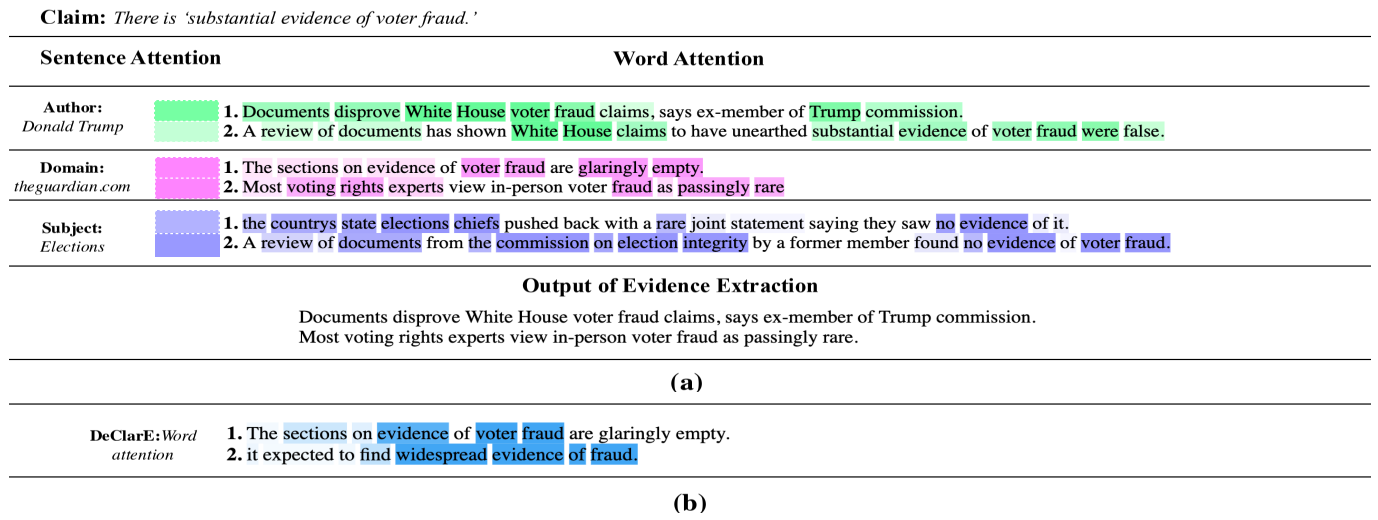



Figure 5: Example 2: Comparison of SADHAN evidence extraction with DeClarE for the claim "There is substantial evidence of voter fraud"

the articles related to him.

In similar fashion in domain model in Figure 4(a) second row, domain based attention is able to capture words 'grossly inaccurate' and 'falsehood' and in Figure 5(a) second row, words like 'glaringly empty' and 'passingly rare', which are otherwise not possible to get attended with just claim only attention. As many articles from same domain, might be written by the same columnist or journalist and hence domain attention tries to capture their writing style and usage of specific phrases or words.

In case of subject model in Figure 4(a), subject based attention learns to attend words and sentences which are related to the subject. As we can see 'Taxes' as subject captures words 'over-taxed' and 'income tax ' etc but also at the sentences level, it is able to capture very interesting sentences like sentence 2. In case of DeClarE model however, the model is unable to attend the most important words and sentences except few, like in sentence 4, though it attends words like 'highly taxed nation' etc but fails to attend word 'falsehood' as we can see in Figure 4(b). As DeClarE model doesn't have sentence level attention, it's therefore not able to use the evidence provided by sentence 4 to decide the appropriate label.

Finally, we show a snippet extracted by our evidence extraction algorithm in Figure 4(a) fourth row and 5(a) fourth row. The value of K is 5 in Figure 4(a) and 2 in 5(a), which means snippet contains top 5 sentences and top 2 sentences based on our evidence extraction method. It is evident that such a sentence extraction technique can be really effective in case of extractive text summarization tasks.

# 7 Conclusions and Future Work

In this paper we presented an hierarchical attention mechanism to jointly learn various latent aspect embeddings for news. For example, these latent aspects can be subject, author and domain related to the claim and news articles. This allows us to capture salient vocabulary and complex structure at the document level. Compared to the only claim-text based attention, the attention weights, which are jointly learned guided by both claim text and different latent aspects are more effective for detecting if the

claims are True or False. This is apparent from our experiments conducted on Snopes, Politifact and Fever dataset. We also propose an algorithm to extract evidence snippets supporting or refuting the claim from news articles using attention weights at both the word and sentence level from all three models. We show a t-SNE visualization that the learned embeddings are also good predictors of fake news. We also show examples where the evidence extracted using our latent aspect embeddings are superior to simple word level attention used in DeClarE. In future, we plan to conduct a detailed user study on the informativeness and interpretability of these evidence snippets.

# References

[1] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning." In: February (2019).

[2] **Kai Shu, Suhang Wang, and Huan Liu**. "Beyond News Contents: The Role of Social Context for Fake News Detection." In: *WSDM*. ACM. 2019, pp. 312–320.

[3] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks." In: July 2018, pp. 1980–1989.

[4] **Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum**. "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning." In: *EMNLP*. 2018, pp. 22–32.

[5] **Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein**. "A Stylometric Inquiry into Hyperpartisan and Fake News." In: *ACL*. Vol. 1. 2018, pp. 231–240.

[6] **Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu**. "Neural User Response Generator : Fake News Detection with Collective User Intelligence." In: *Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* (2018), pp. 3834–3840.

[7] **James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal**. "FEVER: a Large-scale Dataset for Fact Extraction and VERification." In: New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074.

[8] **Liang Wu and Huan Liu**. "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate." In: WSDM '18. 2018.

[9] **Wenpeng Yin and Dan Roth**. "TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification." In: *EMNLP*. 2018, pp. 105–114.

[10] **Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang**. "Fake news detection through multi-perspective speaker profiles." In: *IJCNLP*. Vol. 2. 2017, pp. 252–256.

[11] **Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea**. "Automatic detection of fake news." In: *arXiv preprint arXiv:1708.07104* (2017).

[12] **Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum**. "Where the truth lies: Explaining the credibility of emerging claims on the web and social media." In: *WWW*. 2017, pp. 1003–1012.

[13] **Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi**. "Truth of varying shades: Analyzing language in fake news and political fact-checking." In: *EMNLP*. 2017, pp. 2931–2937.

[14] **Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel**. "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task." In: (2017), pp. 1–6. eprint: 1707.03264.

[15] **Natali Ruchansky, Sungyong Seo, and Yan Liu**. "Csi: A hybrid deep model for fake news detection." In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM. 2017, pp. 797–806.

[16] **Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha**. "Detecting Rumors from Microblogs with Recurrent Neural Networks." In: *IJCAI*. 2016, pp. 3818–3824.

[17] **Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit**. "A Decomposable Attention Model for Natural Language Inference." In: (2016). ISSN: 0001-0782. eprint: 1606.01933.

[18] **Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy**. "Hierarchical attention networks for document classification." In: *NAACL: HLT*. 2016, pp. 1480–1489.

[19] **Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier**. "TweetCred: Real-Time Credibility Assessment of Content on Twitter." In: *SocInfo*. 2014, pp. 228–243.

[20] **Yoon Kim**. "Convolutional neural networks for sentence classification." In: *arXiv preprint arXiv:1408.5882* (2014).

[21] **Alex Graves, Santiago Fernández, and Jürgen Schmidhuber**. "Bidirectional LSTM networks for improved phoneme classification and recognition." In: *ANN*. Springer. 2005, pp. 799–804.

# Paper II:
# Fake News Detection using Higher-order User to User Mutual-attention Progression in Propagation Paths

# Fake News Detection using Higher-order User to User Mutual-attention Progression in Propagation Paths

**Rahul Mishra**[1]

[1] University of Stavanger
Stavanger, Norway
[2] Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway

Published in 2020, The 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

**Abstract:**
Social media has become a very prominent source of news consumption. It brings forth multifaceted, multimodal and real-time information on a silver platter for the users. Fake news or rumor mongering on social media is one of the most challenging issues pertaining to present web. Previously, researchers have tried to classify news propagation paths on social media (e.g. Twitter) to detect fake news. However, they do not utilize latent relationships among users efficiently to model the influence of the users with high prestige on the other users, which is a very significant factor in information propagation. In this paper, we propose a novel **Hi**gher-order User to User **M**utual-**a**ttention **P**rogression (**HiMaP**) method to capture the cues related to authority or influence of the users by modelling direct and indirect (multi-hop) influence relationships among each pair of users, present in the propagation sequence. The proposed higher order attention trick is a novel contribution which can also be very effective in case of transformer architectures[17]. Our model not only outperforms the state-of-the-art methods on two publicly available Twitter datasets but also explains the propagation patterns pertaining to fake news by visualizing higher order mutual-attentions.

# 1 Introduction

Social Media platforms have become part and parcel of our daily lives and are also being used as a common ground for discussions and debates. Rumors and fake news on social media platforms have become a common phenomenon, curbing them is a very challenging and daunting task. The spread of a viral news item or a tweet can seriously affect the election outcomes, reputation of some companies or even relationships among countries, therefore prevailing the sanity in such platforms is the need of the hour. Several machine learning based solutions are investigated in the literature to detect and mitigate the effects of fake news.

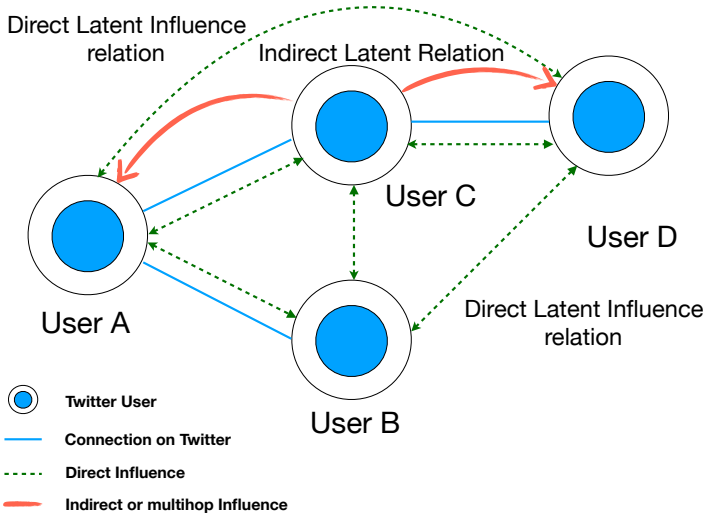



Figure 1: Latent Influence Relationships among users

Previous studies in the literature have used many different facets and aspects related to news items for fake news detection such as the content of the news, source of the news, user response on news and propagation patterns on social media platforms. The news content oriented solutions use handcrafted text or linguistic features to learn a classifier [14, 33, 13], some other works use deep learning techniques instead, to automatically learn the representative features [20, 15]. Recently, neural attention based techniques are also proposed by researchers to detect the misinformation

and they also extract evidences pertaining to classifier's decision as a bi-product [9, 3]. Some of the other interesting works use only temporal propagation patterns of the news items on social media to detect the rumors [7, 8]. The advantage of the propagation patterns based methods over news content and user response oriented methods is that they do not rely on user comments and replies as at an early stage of news propagation, these features are not available readily. However, there are some limitations of these approaches, firstly, they use temporal user characteristics such as number of user followers and followings, the numbers of tweets and retweets posted, which requires tedious feature engineering and transformations. Secondly, they do not model the influence or affinity relationships among the users, which is a key factor in information propagation on social media platforms.

We propose a novel Higher Order User to User Mutual-attention Progression (HiMaP) method to address the limitations of existing methods. Rather than using handcrafted user characteristics features, we use user embeddings, learned via several node embedding techniques. We use user-to-user mutual-attention method to model latent influence relationship among users in propagation paths, which inherently captures the patterns and connotations pertaining to rumor and non-rumor propagation. In figure 1, there are four twitter users $A, B, C$ and $D$ represented as circles and the blue connection lines represent the way they are connected on twitter. Let's assume all of the above mentioned users are the part of a propagation path of a news item $n$. We compute two kinds of latent influence relationship among the users $A, B, C$ and $D$. Firstly, we compute direct user to user influence relationship using mutual-attention such as $A \leftrightarrow B$, $A \leftrightarrow C$, $A \leftrightarrow D$, $B \leftrightarrow C$, $B \leftrightarrow D$ and so on, which are depicted as green dotted connection lines in figure 1. Secondly, we compute indirect user to user influence relationship using higher order mutual-attention progression method such as $A \leftarrow C \rightarrow D$, which is depicted as red stroke lines in figure 1.

We use two publicly available twitter datasets for evaluation and analysis, the proposed model outperforms all the baselines and state of the art methods.

In nutshell, major contributions of this paper are:

(1) We are the first to use the User to User mutual-attention in propagation paths to model and capture the latent cues related to authority or influence of the users.

(2) We enhance the User to User mutual-attention by introducing a novel High Order Mutual-attention Progression method (HiMaP) to model multi-hop latent relationships among the users.

(3) Contrary to previous works, we use both the follower and the retweet networks to learn user embeddings rather than representing users with user characteristics vector.

(4) We achieve significant gains over state-of-the-art models in terms of accuracy, on two publicly available twitter datasets.

(5) We visualize and analyse the attention weights to check the efficacy of the attention mechanism.

# 2 Related Work

We can categorize the previous works related to fake news detection into three major categories based on what features they utilize, 1. news content and linguistic feature oriented, 2. user action on news oriented and 3. social context oriented. The first category of works use text content of the news items, extract several linguistic and statistical features and learn a classifier to detect whether or not it is a rumor [33, 13, 34, 29]. The authors of [14] use language stylistic feature and source credibility features to model the credibility of web claims. The second category of works use user actions on news, such as sentiments, comments, replies and disapproval. In [5], authors use Bayesian network model (probabilistic graphical model with Gibbs sampling) to capture the conditional dependencies among the truthfulness of news, the users' opinions, and the users' credibility. Authors of [10] propose a CNN based model with a user response generator, which learns to generate a synthetic user response to a news article text from historical user responses, which is used as a user action feature in fake news detection.

The third kind of works utilize social context in terms of user profile, social network features and news propagation paths [25, 4, 16]. The authors of

[7] transform the news propagation into a multivariate time series of user characteristics and learn a classifier with concatenated representation of RNN-Based and CNN-Based propagation path representations. Authors in [8] use tree structured neural networks to represent propagation paths, recursive nature of their model effectively captures the tree features of the propagation trees. Authors of [12] propose a new kind of community preserving user embedding method and convert the news propagation tree structures into a temporal sequence and then apply RNN with early stopping for classification.

In contrast to these existing works, HiMaP uses a novel mutual-attention progression model to learn better propagation path representation along with RNN based sequence encoder, which contains cues related to both compositional aspects and latent influence aspects of the propagation sequence.

# 3 Problem Definition and Proposed Model

## 3.1 Problem Definition

Given a news item $n \in N$, along with its propagation path on twitter as $u_1 \rightarrow u_2 \rightarrow u_3....u_{m-1} \rightarrow u_m$, where $u_1$ is the user, who has posted the original tweet about news $n$. $u_m$ is the last user in the sequence, who has retweeted the same tweet. The goal is to classify the news as one of these classes:"True (T)" or "False (F)" or "Unverified (U)" or "Debunking (D)".

## 3.2 Retweet Propagation Path Representation

We represent the propagation paths of the news by sequence of users pertaining to the original (source) tweets and re-tweets as variable length multivariate time series, very similar to [7]. From the original propagation trees, we create a flattened representation of the tree as a multivariate time series comprising user embeddings and timestamp. For a news item $n_i$, propagation sequence $Prop(n_i)$ can be defined as:

$$Prop(n_i) =< (f(u_1), t_0).....(f(u_m), t_m) > \tag{5.1}$$

Where $(f(u_i), t_i)$ represents $i^{th}$ user, $i^{th}$ timestamp and $m$ is the length of propagation sequence. In contrast to [7], we represent each user with learned user embeddings by applying suitable node embedding methods to follower network and retweet network rather than representing users as their characteristics vectors, as depicted in figure 2. Usage of node embeddings instead of characteristics vectors, not only saves the time required to crawl the characteristic features for each user but also does not require any feature engineering.

## 3.3 Learned User Embeddings

We use unsupervised network representation learning methods to learn user (node) embeddings from both the follower network and the retweet network and we combine the corresponding embeddings for each user by concatenating them. Given a follower graph $F = (V, E)$ and a retweet graph $R = (V', E')$, we compute user embedding $f(v)$ for each user $v$ by concatenating user embedding learned from follower network $f_F(v) \in \mathbb{R}^d$ and the user embedding learned from retweet network $f_R(v) \in \mathbb{R}^d$ as:

$$f(v) = f_F(v) \parallel f_R(v) \tag{5.2}$$

Provided $v \in V$ and $v \in V'$. Specifically we experiment with DeepWalk [30], Node2vec [19], Line [26] and APP[18] node embedding methods and select the best performing embedding technique.

- **DeepWalk**: It is a uniform random walk simulation based method, which uses SkipGram with hierarchical softmax as optimizer and objective function as follows:

$$min_\phi - \log P(\{v_{i-w}, ..., v_{i-1}, v_{i+1}, ..., v_{i+w}\}|\phi_i) \tag{5.3}$$

- **Node2vec**: It is a breadth first (BFS) and depth first search (DFS) based method, which uses SkipGram with negative sampling and objective function as follows:

$$max_f \sum_{u \in V} [-\log Z_u + \sum_{n_i \in N_s(u)} f(n_i).f(u)] \tag{5.4}$$

- **APP**: It is a Personalized PageRank Context based method, which uses negative sampling and objective function as follows:

$$\log \sigma(\vec{s_u}.\vec{t_v}) + k.E_{tn}P_D[\log \sigma(\vec{s_u}.\vec{t_n})] \tag{5.5}$$

- **Line**: It is a Adjacency matrix based method, which models neighbourhood proximity using negative sampling and objective function as follows:

$$O1 = \sum_{(i,j) \in E} w_{ij} \log p_1(v_i, v_j) \tag{5.6}$$

For more details about these node embedding methods, please refer to respective papers.

## 3.4 LSTM based Propagation Path Sequence Encoder

We use Long short term memory unit (LSTM) [36] to encode the propagation path sequence $f(u_1) \rightarrow f(u_2) \rightarrow f(u_3)....f(u_{m-1}) \rightarrow f(u_m)$ represented as a sequence of learned user embeddings. At a particular time-step $t$, the current hidden state $h_t$ is computed using standard LSTM equations as follows:

$$\begin{aligned} f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i.[h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C.[h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t), \end{aligned} \tag{5.7}$$

Where, $h_{t-1}$ is previous hidden state and $x_t$ is the current input from input propagation path sequence. We use last hidden state as a compositional representation of propagation path sequence $R_C$, where

$$R_C = h_m \tag{5.8}$$

## 3.5 User to User Mutual-attention

We explain the user to user mutual-attention in detail in this section, as pictorially represented in the figure 2. Neural attention [28] mechanisms are proven to be very effective in many NLP [17, 22, 2] and computer vision applications [27, 23, 24]. The key idea behind the neural attention is to select the important words or sentences in NLP applications and to gauge the crucial areas or blocks in images in typical computer vision applications. Many of the previous works use attention mechanisms to detect fake news[9, 3] and to do fact checking[21]. It is well studied fact that influential users play a very crucial role in information diffusion on social media platforms[6, 32] on the other hand it's very hard to quantify the influence and it's penetration in a real world social network. Interpersonal relationships among users are the key factor in determining the influence [35]. We are the first to propose a User to User mutual-attention method to model the influence among the users. Previously, researchers have used mutual-attention mechanism in case of word to word mutual-attention within a sentence to model intra-relationships among words, present in same sentence [11]. Given a propagation path of a news item $n$, in terms of sequence of learned user embeddings as $f(u_1) \rightarrow f(u_2) \rightarrow f(u_3)....f(u_{m-1}) \rightarrow f(u_m)$, in the first step we model the relationship among each pair of users present in the propagation path. In very similar fashion to[11], We use a dense layer to project the concatenation of each user embedding pair into a scalar score:

$$S_{ij} = W_{cat}([f(u_i); f(u_j)]) + b_{cat} \tag{5.9}$$

Where $W_{cat} \in \mathbb{R}^{2d \times 1}$ is a weight matrix, $b_{cat} \in \mathbb{R}$ is bias term and $S_{ij}$ is the latent affinity between users $u_i$ and $u_j$. Score matrix is $\underset{m \times m}{\mathrm{S}}$ is a square

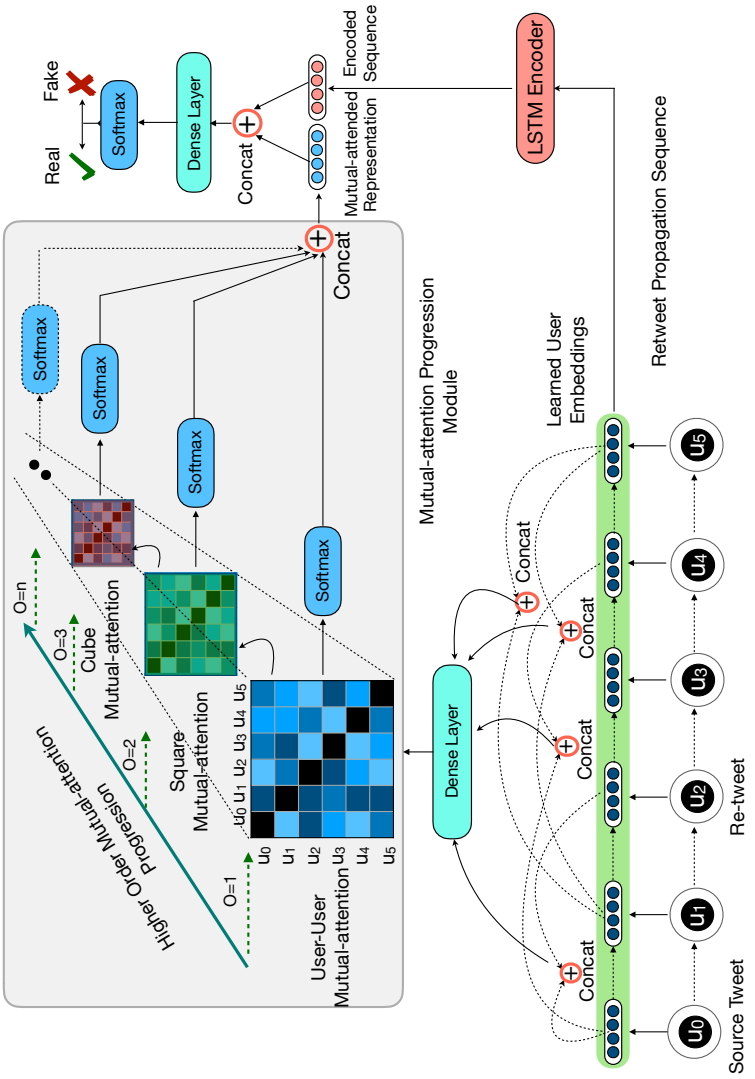



Figure 2: high-level architecture of HiMaP with Higher Order Mutual-attention Progression

matrix. To compute mutual-attention scores, we can consider two options, either we can apply row-wise max-pooling or row-wise avg-pooling.

$$\begin{aligned} A_C &= Softmax(\underset{row}{max}S) \\ &\textit{or} \\ A_C &= Softmax(\underset{row}{avg}S) \end{aligned} \tag{5.10}$$

Where $A_C \in \mathbb{R}^m$ is the learned attention weight vector. Finally, user to user mutuallly-attended representation $R_A$ of the propagation path can be computed as:

$$R_A = \sum_{i=1}^{m} f(u_i) A_{Ci} \tag{5.11}$$

## 3.6 Multi-hop Latent Relationships

As of now in user to user mutual-attention, we only consider influence in terms of attention between each pair of users present in the propagation path individually, which only models relationship between two users at a time regardless of the presence of other users in the sequence. In a real world social network scenario, in some of the cases users only trust and subsequently retweet the content if and only if some particular combination of users have already posted or retweeted the content in their social network fraternity. We call these scenarios as multi-hop latent relationships, in which influence depends on a group of users rather than a single user. We can not capture cues related to such multi-hop latent relationships with the first order user to user mutual-attention described earlier. We propose a novel higher order mutual-attention progression method to deal with it.

## 3.7 Higher Order mutual-attention Progression

The proposed Higher Order attention progression method is a novel theoretical contribution in the neural attention domain. The intuition behind mutual-attention progression is fairly simple. In the equation 5.9, values in

the score matrix $\underset{m\times m}{\mathrm{S}}$ represent the direct influence relationships between each possible user pairs in the propagation path. Now let's consider a matrix $\underset{m\times m}{\mathrm{S}^2}$ which is computed as:

$$\underset{m\times m}{\mathrm{S}^2} = \underset{m\times m}{\mathrm{S}} \times \underset{m\times m}{\mathrm{S}} \tag{5.12}$$

Each value in the matrix $S^2$ represents the indirect influence or affinity between two given users in the propagation path sequence.

$$S^2_{i,j} = \sum_k S_{i,k} \times S_{k,j} \tag{5.13}$$

This represents the influence between pair of users $i$ and $j$, encompassing all other users. In the similar fashion we can compute more higher order influence matrices.

$$\begin{aligned} \underset{m\times m}{\mathrm{S}^3} &= \underset{m\times m}{\mathrm{S}^2} \times \underset{m\times m}{\mathrm{S}} \\ \underset{m\times m}{\mathrm{S}^4} &= \underset{m\times m}{\mathrm{S}^3} \times \underset{m\times m}{\mathrm{S}} \end{aligned} \tag{5.14}$$

Now to compute attention scores, we use row-wise max pooling similar to equation 4 as:

$$\begin{aligned} A'_{Co} &= Softmax(\underset{row}{max} S^2) \\ A''_{Co} &= Softmax(\underset{row}{max} S^3) \\ A'''_{Co} &= Softmax(\underset{row}{max} S^4) \end{aligned} \tag{5.15}$$

Where $A'_{Co}$, $A''_{Co}$ and $A'''_{Co}$ are the learned attention weight vectors from second order, third order and fourth order of mutual-attention progression. Finally, higher order user to user mutuallly-attended representations $R'_C$, $R''_C$, $R'''_C$ etc can be computed as:

$$
\begin{aligned}
R'_A &= \sum_{i=1}^{m} f(u_{i)A'_{Coi}} \\
R''_A &= \sum_{i=1}^{m} f(u_{i)A''_{Coi}} \\
R'''_A &= \sum_{i=1}^{m} f(u_{i)A'''_{Coi}}
\end{aligned}
\tag{5.16}
$$

## 3.8 Prediction Layer

At the prediction stage, we have two kinds of representations of the propagation path sequence, a representation encoded by LSTM based encoder as $R_C$ and the representations computed using higher order mutual-attention progression as $R'_A$, $R''_A$, $R'''_A$, $R''''_A$ and so on, representing first order, second order, third order and fourth order mutual-attention representations. We compute the cumulative representation from the all higher order mutual-attention progression representations by concatenating them.

$$
{R^f}_A = R'_A \parallel R''_A \parallel R'''_A \parallel R'''_A \tag{5.17}
$$

We learn a joint representation of ${R^f}_A$ and $R_C$ using a non linear transformation layer.

$$
R = ReLU(W_p([{R^f}_A, R_C] + b_p) \tag{5.18}
$$

At the end, we use a Softmax layer for the classification.

$$
\hat{y} = Softmax(W_{cl}R + b_{cl}) \tag{5.19}
$$

## 3.9 Optimization

We use standard Softmax cross-entropy with logits as loss function to train our model.

$$L = -\sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}} \tag{5.20}$$

where $L$ is the cost function to be minimized, $y_i$ is class label of $x_i$.

# 4 Experimental Setup

## 4.1 Research Questions

We conduct the experiments with objective to find answers to following research questions:

(1) **RQ1**: Is the proposed user to user mutual-attention mechanism useful for the fake news classification?

(2) **RQ2**: Does the higher order mutual-attention progression provide useful new and uncovered cues or patterns?

(3) **RQ3**: Does the proposed models outperform the state of the art models?

## 4.2 Datasets

Table 5.1: Dataset Statistics

| **Statistics** | **Twitter15** | **Twitter16** |
|---|---|---|
| News items | 1490 | 818 |
| True news | 374 | 205 |
| fakenews | 370 | 205 |
| Unverified | 374 | 203 |
| Debunking | 372 | 205 |
| Users | 276663 | 173487 |
| Posts | 331612 | 204820 |
| Followers | 359385237 | 225359613 |
| Followings | 398394720 | 249821280 |

We use two publicly available twitter datasets[5] [1] called Twitter15 and Twitter16 for the evaluation. Twitter15 dataset contains 1490 news stories and Twitter16 dataset contains 818 news stories. In table 5.1, some statistics related to datasets are shown, for more details of the dataset statistics please refer to [1]. We use Twitter API[6], to crawl the user followers and following as these are not present in datasets.

## 4.3 Baselines and variants of proposed model

We compare the proposed model with several baseline and state of the art works.

- **DTC:** [33] This work uses hand crafted text and other statistical features with a decision tree classifier to asses credibility of tweets.
- **SVM-RBF:**[31] This work uses a radial basis function kernel based SVM model to classify news as rumor or non-rumor.
- **SVM-TS:**[25] In this paper, authors create a time series of news characteristics and classify using a SVM model.
- **GRU-RNN:**[**Ma16**] A gated recurrent unit based model which learns propositional representation of rumors and non-rumors.
- **TD-RvNN:**[8] This work utilizes tree-structured neural networks for rumor representation learning and classification.
- **PPC-RNN+CNN:**[7] In this method, authors propose a multivariate time series representation of news propagation and use combination of GRU and CNN models for classification.

We compare results of above mentioned models with three variants of our HiMaP model.

- **HiMaP-FO:** This is the HiMaP model with first order mutual-attention, where order $O = 1$.

[5] `https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0`

[6] `https://dev.twitter.com/rest/public`

Table 5.2: Comparison of proposed model with various state of the art baseline models for twitter15 and twitter16 datasets. HiMaP-HO is statistically significant ($p - value = 2.75e^{-3}, 2.03e^{-4}$ for Twitter15 and Twitter16 using pairwise student's t-test)

| Twitter15 | | | | | |
|---|---|---|---|---|---|
| **Model** | **Acc.** | **T F1** | **F F1** | **U F1** | **D F1** |
| DTC | 0.442 | 0.731 | 0.351 | 0.320 | 0.423 |
| SVM-RBF | 0.326 | 0.442 | 0.048 | 0.241 | 0.273 |
| SVM-TS | 0.548 | 0.773 | 0.488 | 0.403 | 0.479 |
| GRU-RNN | 0.641 | 0.684 | 0.634 | 0.688 | 0.571 |
| TD-RvNN | 0.723 | 0.682 | 0.758 | 0.821 | 0.654 |
| PPC-RNN+CNN | 0.842 | 0.811 | 0.875 | 0.790 | 0.818 |
| HiMaP-FO | 0.863 | 0.822 | 0.901 | 0.814 | 0.826 |
| HiMaP-HO | 0.869 | 0.831 | 0.889 | **0.835** | 0.828 |
| HiMaP-HO+Text | **0.880** | **0.837** | **0.917** | 0.834 | **0.830** |
| Twitter16 | | | | | |
| **Model** | **Acc.** | **T F1** | **F F1** | **U F1** | **D F1** |
| DTC | 0.462 | 0.742 | 0.335 | 0.337 | 0.434 |
| SVM-RBF | 0.331 | 0.442 | 0.085 | 0.251 | 0.219 |
| SVM-TS | 0.572 | 0.809 | 0.469 | 0.421 | 0.494 |
| GRU-RNN | 0.649 | 0.691 | 0.628 | 0.719 | 0.592 |
| TD-RvNN | 0.743 | 0.705 | 0.772 | 0.842 | 0.671 |
| PPC-RNN+CNN | 0.863 | 0.826 | 0.883 | 0.810 | 0.824 |
| HiMaP-FO | 0.882 | 0.842 | 0.936 | 0.832 | 0.843 |
| HiMaP-HO | 0.890 | 0.844 | 0.921 | **0.858** | **0.857** |
| HiMaP-HO+Text | **0.913** | **0.849** | **0.939** | 0.854 | 0.854 |

- **HiMaP-HO:** This is the HiMaP model with higher order mutual-attention, where order $O \geq 2$.
- **HiMaP-HO+Text:** This is the HiMaP model with higher order mutual-attention, where order $O \geq 2$ and we also use an LSTM sequence encoder to encode original news text along with propagation path sequence.

Table 5.3: Performance of mutual-attention Progression method with higher orders

| Twitter15 | | Twitter16 | |
|---|---|---|---|
| **Model** | **Acc.** | **Model** | **Acc.** |
| HiMaP-FO (O=1) | 0.8631 | HiMaP-FO (O=1) | 0.8828 |
| HiMaP-HO (O=2) | 0.8663 | HiMaP-HO (O=2) | 0.8891 |
| HiMaP-HO (O=3) | 0.8696 | HiMaP-HO (O=3) | 0.8901 |
| HiMaP-HO (O=4) | 0.8696 | HiMaP-HO (O=4) | **0.8908** |
| HiMaP-HO (O=5) | **0.8697** | HiMaP-HO (O=5) | **0.8908** |

Table 5.4: Performance of HiMaP with different node embedding methods

| Twitter15 | | Twitter16 | |
|---|---|---|---|
| **Model** | **Acc.** | **Model** | **Acc.** |
| DeepWalk | 0.825 | DeepWalk | 0.850 |
| Node2Vec | 0.846 | Node2Vec | 0.867 |
| APP | 0.861 | APP | 0.889 |
| Line | **0.869** | Line | **0.890** |

## 4.4 HiMaP Implementation

We use TensorFlow framework to implement [7] our proposed models. We compute overall accuracy and per class F1 scores as performance metrics for evaluation and comparison with the state of the art methods. We use softmax cross entropy with logits as the loss function, learning rate of 0.003 and size of hidden states LSTM units are kept as 100. We tune all the parameters using random search. We use $50$ epochs for each model and use dropout regularization ($keepprob = 0.2$) and early stopping if validation loss does not change for more than 10 epochs. We observe that the sequence length of $35$ gives the optimal performance in both the twitter datasets.

The user embeddings are learned using various node embeddings methods namely, DeepWalk [30], Node2vec [19], Line [26] and APP[18]. For all the node embedding methods, we use prescribed parameters and em-

[7] `https://github.com/rahulOmishra/HiMaP`

bdedding size as 100. From the retweet networks (trees), we extract all the unique edges and nodes. We assign the weight for each edge as the number of times it occurs in our network. Similarly in case of follower network, we extract all the unique edges and nodes and use node embedding methods to train the node embeddings for each node involved in the network. In case of HiMaP-FO+Text model, we use pretrained GloVe embeddings of 100 dimensions as word embeddings.

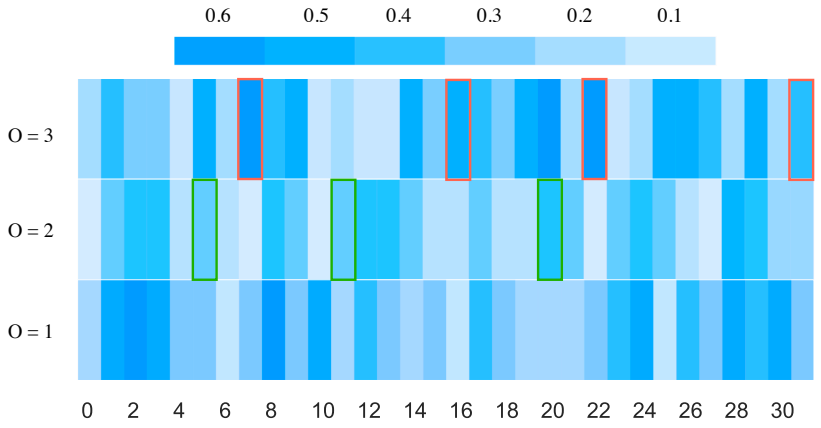



Figure 3: Normalized mutual-attention weight visualization

# 5 Experimental Results and Analysis

In this section, we evaluate the proposed model and analyse the significance of attention mechanism.

## 5.1 Results for Twitter15 and Twitter16 datasets

In Table 5.2, we present the comparison of performance of the proposed model with several baselines and state-of-the-art methods. We can observe

that even the basic HiMaP model (HiMaP-FO, where $O = 1$) outperforms all the baselines and state of the art models. Among the baseline methods, we notice that RNN based methods are more effective than other methods. An intuitive explanation for this trend can be the capability of RNN models to easily learn the compositional aspects of news content in GRU-RNN and news propagation sequence in PPC-RNN+CNN and TD-RvNN, without any or with minimal feature engineering. On the other hand, PPC-RNN+CNN outperforms TD-RvNN as they use CNN to capture the local variations within a propagation sequence.

The basic HiMaP model (HiMaP-FO, where $O = 1$) performs better than both the state of the arts (PPC-RNN+CNN and TD-RvNN) as it not only uses LSTM based sequence encoder to capture the compositional aspects but also utilizes user-user mutuallly-attended representations of propagation path, which inherently holds the latent cues and patterns pertaining to influence relationships among users. Therefore we can conclude that research questions **RQ1** and **RQ3** are satisfied. The HiMaP-HO (where $O \geq 2$) model outperforms HiMaP-FO model as it uses mutual-attention progression of higher order, which captures indirect relationships among all pairs of users present in the propagation path sequence.

The HiMaP-HO+Text model uses original news text also with propagation path, which provides additional topical and semantic cues related news text and outperforms all the other models.

## 5.2 Analysis of HiMaP with Higher Order mutual attention

In table 5.3 we show comparison of HiMaP model with different values of mutual-attention order $O$. We can observe that performance of HiMaP in terms of accuracy improves with increase in mutual-attention order, this means research question **RQ2** is answered partially as we know now that higher order mutual-attention is useful. However, we notice that after $O = 3$, accuracy starts to saturate for higher orders in both the datasets, therefore we can tune the parameter $O$ and omit computation of higher order mutual-attention. In table 5.2, where we compare HiMaP with baselines and state of the art models, we use HiMaP results with third

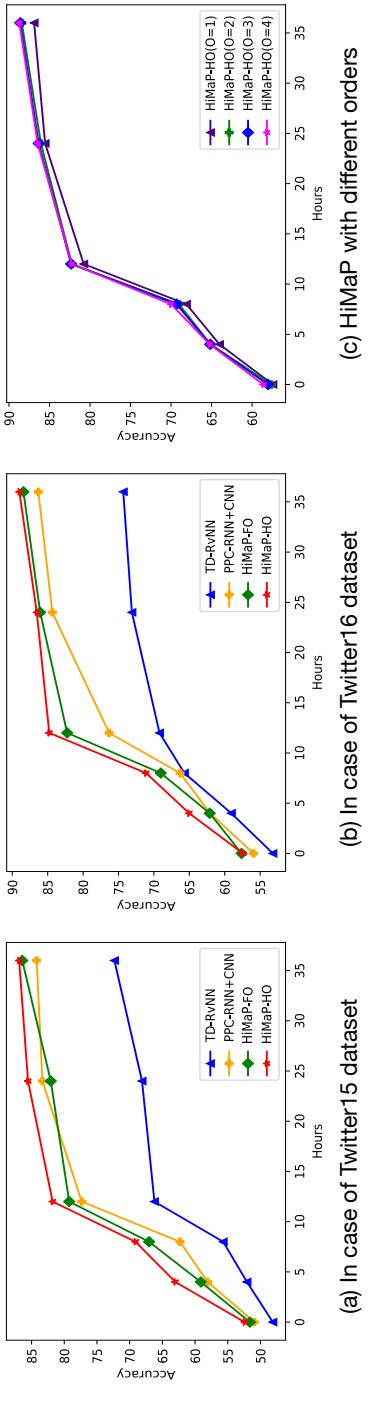



Figure 4: Comparison of HiMaP models with state-of-the-art models in terms of early detection accuracy at different timestamps of the news propagation, depicted in plots (a) and (c). Comparison of HiMaP models with different mutual-attention levels in terms of early detection accuracy at different timestamps of the news propagation, depicted in plot (c).

order of mutual-attention progression where $O = 3$.

## 5.3 Analysis of HiMaP with different node embedding methods

In table 5.4, we show the effect of using different node embedding methods to learn user embeddings. We observe that the Line method outperforms all the other node embedding methods in both the datasets. In table 5.2, where we compare HiMaP with baselines and state of the art models, we use HiMaP results with Line embeddings. The reason behind the better performance of the Line method can be the suitability of the Line method for graphs with low clustering coefficient and transitivity and we observe for both the Twitter datasets (twitter15 and twitter16), the values of clustering coefficient and transitivity are low.

## 5.4 Comparison of Early Detection Accuracy

In Figure 4, we compare the early detection performance of the proposed models with the state-of-the-art models. We plot the overall accuracy vs elapsed time since the original tweet is posted. We can observe in Figure $4(a)$ and $4(b)$ that for both the twitter15 and twitter16 datasets, HiMaP-HO outperforms state-of-the-art models at each time step. The better performance of HiMaP can be attributed to the learning of additional and useful propagation patterns due to higher-order mutual-attention progression method. In Figure 4, we also compare the early detection performance of HiMaP models with different values of order $O$. We can observe in $4(c)$ that there is significant improvement form $O = 2$ to $O = 3$ but there is not much significant improvement above third order ($O \geq 3$).

## 5.5 Mutual-attention Visualization and Analysis

In this section, we explain the visualization of attention weights from three levels of mutual-attention progression, for a propagation sequence of a anecdotal news example. In figure 3, there are $32$ users in the propagation path of news item. Each strip in figure 3, represents a different order of mutual-attention, first strip is the depiction of attention weights from first order mutual-attention, where $O = 1$ and similarly second and third strip

depict weights from second and third order mutual-attention weights for the same propagation sequence. The depth of the colors in the rectangles in each strip represents the distribution of attention weights among users present in propagation path. We do not reveal the identity of users for sake of privacy and twitter's policy. We observe that in the first order mutual-attention, users with high number of followers get more attention weights. In contrast to first order mutual-attention, in the second and third order, some of the users with less followers and prestige also get higher attention weights (highlighted rectangles with green and red borders). We also notice that beyond third order mutual-attention $O = 3$, there are no significant changes in the attention pattern. We conclude that higher order mutual-attention captures new, uncovered and significant latent patterns and hence research question **RQ2** is satisfied.

# 6 Conclusions

In this paper, we propose a novel user to user mutual-attention progression method to model influence relationships among users, present on news propagation path to detect fake news. This method allows us to capture both the direct and indirect (multi-hop) relationships between each pair of users. Experiment with two publicly available twitter datasets, shows the effectiveness of our model, compared to state-of-the-art models, in terms of early detection and overall accuracy. We also notice that higher-order mutual-attention progression method captures useful new, uncovered patterns and provides the classifier with the cues pertaining to propagation of true or fake news. The proposed attention progression trick can also be useful in other application scenarios such as in case of word to word attention in sentences.

In future, we plan to conduct an experiment, related to evidence extraction, using mutual-attention weights from different levels of higher order mutual-attention. Effectiveness of the Higher-order Attention trick can also be utilized with recent transformer architectures.

# References

[1] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning." In: ACL'17, pp. 708–717.

[2] **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[3] **Rahul Mishra and Vinay Setty**. "SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection." In: ICTIR '19. Santa Clara, CA, USA, 2019, pp. 197–204. ISBN: 9781450368810.

[4] **Kai Shu, Suhang Wang, and Huan Liu**. "Beyond News Contents: The Role of Social Context for Fake News Detection." In: *WSDM*. ACM. 2019, pp. 312–320.

[5] **Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu**. "Unsupervised Fake News Detection on Social Media: A Generative Approach." In: AAAI '19. Feb. 2019.

[6] **Sambaran Bandyopadhyay, Ramasuri Narayanam, and M. Narasimha Murty**. "A Generic Axiomatic Characterization for Measuring Influence in Social Networks." In: 2018, pp. 2606–2611.

[7] **Yang Liu and Yi-fang Brook Wu**. "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks." In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018), pp. 354–361.

[8] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks." In: July 2018, pp. 1980–1989.

[9] **Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum**. "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning." In: *EMNLP*. 2018, pp. 22–32.

[10] **Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu**. "Neural User Response Generator: Fake News Detection with Collective User Intelligence." In: IJCAI '18. 2018, pp. 3834–3840.

[11] **Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su**. "Reasoning with sarcasm by reading in-between." In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2018), pp. 1010–1020.

[12] **Liang Wu and Huan Liu**. "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate." In: WSDM '18. 2018.

[13] **Sejeong Kwon, Meeyoung Cha, and Kyomin Jung**. "Rumor Detection over Varying Time Windows." In: vol. 12. Jan. 2017, e0168344. DOI: 10.1371/journal.pone.0168344.

[14] **Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum**. "Where the truth lies: Explaining the credibility of emerging claims on the web and social media." In: *WWW*. 2017, pp. 1003–1012.

[15] **Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi**. "Truth of varying shades: Analyzing language in fake news and political fact-checking." In: *EMNLP*. 2017, pp. 2931–2937.

[16] **Natali Ruchansky, Sungyong Seo, and Yan Liu**. "Csi: A hybrid deep model for fake news detection." In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM. 2017, pp. 797–806.

[17] **Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin**. "Attention Is All You Need." In: abs/1706.03762 (2017).

[18] **Chang Zhou, Yuqiong Liu, Xiaofei Liu, Zhongyi Liu, and Jun Gao**. "Scalable Graph Embedding for Asymmetric Proximity." In: AAAI'17. 2017.

[19] **Aditya Grover and Jure Leskovec**. "Node2vec: Scalable Feature Learning for Networks." In: KDD '16. 2016, pp. 855–864.

[20] **Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha**. "Detecting Rumors from Microblogs with Recurrent Neural Networks." In: *IJCAI*. 2016, pp. 3818–3824.

[21] **Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit**. "A Decomposable Attention Model for Natural Language Inference." In: (2016). ISSN: 0001-0782. eprint: 1606.01933.

[22] **Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy**. "Hierarchical Attention Networks for Document Classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–1489. DOI: 10.18653/v1/N16-1174.

[23] **Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell**. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[24] **Andrej Karpathy and Li Fei-Fei**. "Deep Visual-Semantic Alignments for Generating Image Descriptions." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[25] **Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong**. "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites." In: CIKM '15. Melbourne, Australia, 2015, pp. 1751–1754.

[26] **Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei**. "LINE: Large-Scale Information Network Embedding." In: WWW '15. 2015.

[27] **Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio**. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2015, pp. 2048–2057.

[28] **Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio**. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014. arXiv: 1409.0473 [cs.CL].

[29] **Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier**. "TweetCred: Real-Time Credibility Assessment of Content on Twitter." In: *SocInfo*. 2014, pp. 228–243.

[30] **Bryan Perozzi, Rami Al-Rfou, and Steven Skiena**. "DeepWalk: Online Learning of Social Representations." In: KDD '14. 2014, pp. 701–710.

[31] **Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang**. "Automatic Detection of Rumor on Sina Weibo." In: MDS '12. Beijing, China, 2012.

[32] **Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts**. "Everyone's an Influencer: Quantifying Influence on Twitter." In: WSDM '11. Hong Kong, China, 2011, pp. 65–74. ISBN: 9781450304931.

[33] **Carlos Castillo, Marcelo Mendoza, and Barbara Poblete**. "Information credibility on twitter." In: *WWW*. 2011.

[34] **Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei**. "Rumor Has It: Identifying Misinformation in Microblogs." In: EMNLP '11. 2011. ISBN: 9781937284114.

[35] **Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi**. "Measuring User Influence in Twitter: The Million Follower Fallacy." In: *ICWSM*. 2010.

[36] **Sepp Hochreiter and Jürgen Schmidhuber**. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

# Paper III: Generating Fact Checking Summaries for Web Claims

# Generating Fact Checking Summaries for Web Claims

**Rahul Mishra**[1], **Dhruv Gupta**[2], **Markus Leippold**[3]

[1] Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway
[2] Max Planck Institute for Informatics, Germany
[3] University of Zurich, Switzerland

Published in 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text.

**Abstract:**
We present SUMO, a neural attention-based approach that learns to establish the correctness of textual claims based on evidence in the form of text documents (e.g., news articles or Web documents). SUMO further generates an extractive summary by presenting a diversified set of sentences from the documents that explain its decision on the correctness of the textual claim. Prior approaches to address the problem of fact checking and evidence extraction have relied on simple concatenation of claim and document word embeddings as an input to claim driven attention weight computation. This is done so as to extract salient words and sentences from the documents that help establish the correctness of the claim. However, this design of claim-driven attention does not capture the contextual information in documents properly. We improve on the prior art by using improved claim and title guided hierarchical attention to model effective contextual cues. We show the efficacy of our approach on datasets concerning political, healthcare, and environmental issues.

# 1 Introduction

Most of the information consumed by the world is in the form of digital news, blogs, and social media posts available on the Web. However, most of this information is written in the absence of facts and evidences. Our ever-increasing reliance on information from the Web is becoming a severe problem as we base our personal decisions relating to politics, environment, and health on unverified information available online. For example, consider the following unverified claim on the Web:

*"Smoking may protect against COVID-19."*

A user attempting to verify the correctness of the above claim will often take the following steps: issue keyword queries to search engines for the claim; going through the top reliable news articles; and finally making an informed decision based on the gathered information. Clearly, this approach is laborious, takes time, and is error-prone. In this work, we present SUMO, a neural approach that assists the user in establishing the correctness of claims by automatically generating explainable summaries for fact checking. Example summaries generated by SUMO for couple of Web claims are given in Figure 1.

**Prior approaches** to automatic fact checking rely on predicting the credibility of facts [12], instance detection [6, 10], and fact entailment in supporting documents [16]. The majority of these methods rely on linguistic features [12, 8, 24], social contexts, or user responses [18] and comments. However, these approaches do not help explain the decisions generated by the machine learning models. Recent works such as [1, 3, 7] overcome the explainability gap by extracting snippets from text documents that support or refute the claim. [3, 7] apply claim-based and latent aspect-based attention to model the context of text documents. [3] model latent aspects such as the speaker or author of the claim, topic of the claim, and domains of retrieved Web documents for the claim. We observe in our experiments that in prior works [3, 7], the design of claim guided attention in these methods is not effective and latent aspects such as the topic and speaker of claims are not always available. The snippets extracted by such models are not comprehensive or topically diverse. To overcome these limitations, we propose a novel design of claim and document title driven attention, which better captures the contextual cues in relation to the claim.

| | | | |
|---|---|---|---|
| ***Claim:*** | ***Smoking may protect against COVID-19*** | **Label: False** | **Verdict: False** |
| ***Summary:*** | The current evidence suggests that the severity of COVID is higher among smokers, prevent the health risk linked to the excessive consumption or misuse" of nicotine products by people hoping to protect themselves from COVID-19. Evidence from China, where COVID-19 originated, shows that people who have cardiovascular and respiratory conditions caused by tobacco use, or otherwise, are at higher risk of developing severe COVID-19 symptoms. HO urges researchers, scientists and the media to be cautious about amplifying unproven claims that tobacco or nicotine could reduce the risk of COVID-19. Smoking is also associated with increased development of acute respiratory distress syndrome, a key complication for severe cases of COVID-19. | | |
| ***Claim:*** | ***Deforestation has made humans more vulnerable to pandemics*** | **Label: True** | **Verdict: True** |
| ***Summary:*** | Deforestation can directly increase the likelihood that a pathogen will be transferred from wildlife species to humans through the creation of suitable habitats for vector species. Climate change, including deforestation which drives it, is a key driver of cross-species transmission which is where zoonotic emerging diseases come from . There is a correlation between deforestation and the rise in the spread of infectious diseases affecting humans. Deforestation forces various species into smaller, shared habitats and increases encounters between wildlife and humans. Habitat destruction and fragmentation due to deforestation can also increase the frequency of contact between humans, wildlife species, and the pathogens they carry . This can occur through direct transfer of pathogens from animals to humans or indirectly through cross-species transfer of pathogens from wildlife to domesticated species . Deforestation could be to blame for the rise of infectious diseases like the novel coronavirus. | | |

Figure 1: Example summaries generated by SUMO for unverified claims on the Web.

In addition to this, we propose an approach for generating summaries for fact-checking that are non-redundant and topically diverse.

**Contributions.** Contributions made in this work are as follows. First, we introduce SUMO, a method that improves upon the previously used claim guided attention to model effective contextual representation. Second, we propose a novel attention on top of attention (Atop) method to improve the overall attention effectiveness. Third, we present an approach to generate topically diverse multi-document summaries, which help in explaining the decision SUMO makes for establishing the correctness of claims. Fourth, we provide a novel testbed for the task of fact checking in the domain of climate change and health care.

**Outline.** The outline for the rest of the article is as follows. In Section 2, we describe prior work in relation to our problem setting. In Section 3, we formalize the problem definition and describe our approach, SUMO, to generate explainable summaries for fact checking of textual claims. In Sections 4 and 5, we describe the experimental setup that includes a description of the novel datasets that we make available to the research community and an analysis of the results we have obtained. In Section 6, we present the concluding remarks of our study.

# 2 Related work

We now describe prior work related to our problem setting. First, we describe works that rely only on features derived from documents that support the input textual claim. Second, we describe works that additionally include features derived from social media posts in connection to the claim. Third and finally, we describe works that rely on extracting textual snippets from text documents to explain a model's decision on the claim's correctness.

## 2.1 Content Based Approaches

Prior approaches for fact checking vary from simple machine learning methods such as SVM and decision trees to highly sophisticated deep learning methods. These works largely utilize features that model the

linguistic and stylistic content of the facts to learn a classifier [23, 15, 24, 13]. The key shortcomings of these approaches are as follows. First, classifiers trained on linguistic and stylistic features perform poorly as they can be misguided by the writing style of the false claims, which are deliberately made to look similar to true claims but are factually false. Second, these methods lack in terms of user response and social context pertaining to the claims, which is very helpful in establishing the correctness of facts.

## 2.2 Social Media Based Approaches

Works such as [9, 4, 5] overcome the issue of user feedback by using a combination of content-based and context-based features derived from related social media posts. Specifically, the features derived from social media include propagation patterns of claim related posts on social media and user responses in the form of replies, likes, sentiments, and shares. These methods outperform content-based methods significantly. In [5], the authors propose a probabilistic graphical model for causal mappings among the post's credibility, user's opinions, and user's credibility. In [9], the authors introduce a user response generator based on a deep neural network that leverages the user's past actions such as comments, replies, and posts to generate a synthetic response for new social media posts.

## 2.3 Model Explainability

Explaining a machine learning model's decision is becoming an important problem. This is because modern neural network based methods are increasingly being used as black-boxes. There exist few machine learning models for fact checking that explain this decision via summaries. Related works [3, 7] achieve significant improvement in establishing the credibility of textual claims by using external evidences from the Web. They additionally extract snippets from evidences that explain their model's decision. However, we find that the claim-driven attention design used in these methods is inadequate, and does not capture sufficient context of the documents in relation to the input claim. The snippets extracted by these methods are often redundant and lack topical diversity offered by Web evidences. In contrast, our method enhances the claim-driven attention

mechanism and generates a topically diverse, coherent multi-document summary for explaining the correctness of claims.

# 3 SUMO

We now formally describe the task of fact checking and explain SUMO in detail. SUMO works in two stages. In the first stage, it predicts the correctness of the claim. In the second stage, it generates a topically diverse summary for the claims. As input, we are provided with a Web claim $c \in C$, where $C$ is a collection of Web claims and a pseudo-relevant set of documents $D = \{d_1, d_2, \ldots, d_m\}$, where $m$ is the number of results retrieved for claim $c$. The documents $d \in D$ are retrieved from the Web as potential evidences, using claim $c$ as a query. Each retrieved document $d$ is accompanied by its title $t$ and text body $bd$, i.e. ($d = \langle t, bd \rangle$). We define the representation of each document's body as a collection of $k$ sentences as $bd = \{s_1, s_2, ..., s_k\}$ and each sentence as the collection of $l$ words as $\{w_1, w_2, ..., w_l\} \in \mathbb{W}$, where $\mathbb{W}$ is the overall word vocabulary of the corpus. By $k$ and $l$, we denote the maximum numbers of sentences in a document and the maximum number of words in a sentence, respectively. We use both WORD2VEC and pre-trained GloVe embeddings to obtain the vector representations for each claim, title, and document body. The objective is to classify the claim as either true or false and automatically generate a topically diverse summary pieced together from $D$ for establishing the correctness of the claim.

## 3.1 Predicting Claim Correctness by Neural Attention

We now describe SUMO's neural architecture (see Figure 3) that helps in predicting the correctness of the input claim along with its pseudo-relevant set of documents. The model additionally learns the weights to words and sentences in the document's body that help ascertain the claim's correctness. First, we need to encode the pseudo-relevant documents that support a claim. To this end, as a **sequence encoder**, we use a Gated Recurrent Unit (GRU) to encode the document's body content. Claim and document's title are not encoded using sequence encoder; we explain the method to represent them in detail in upcoming sections.

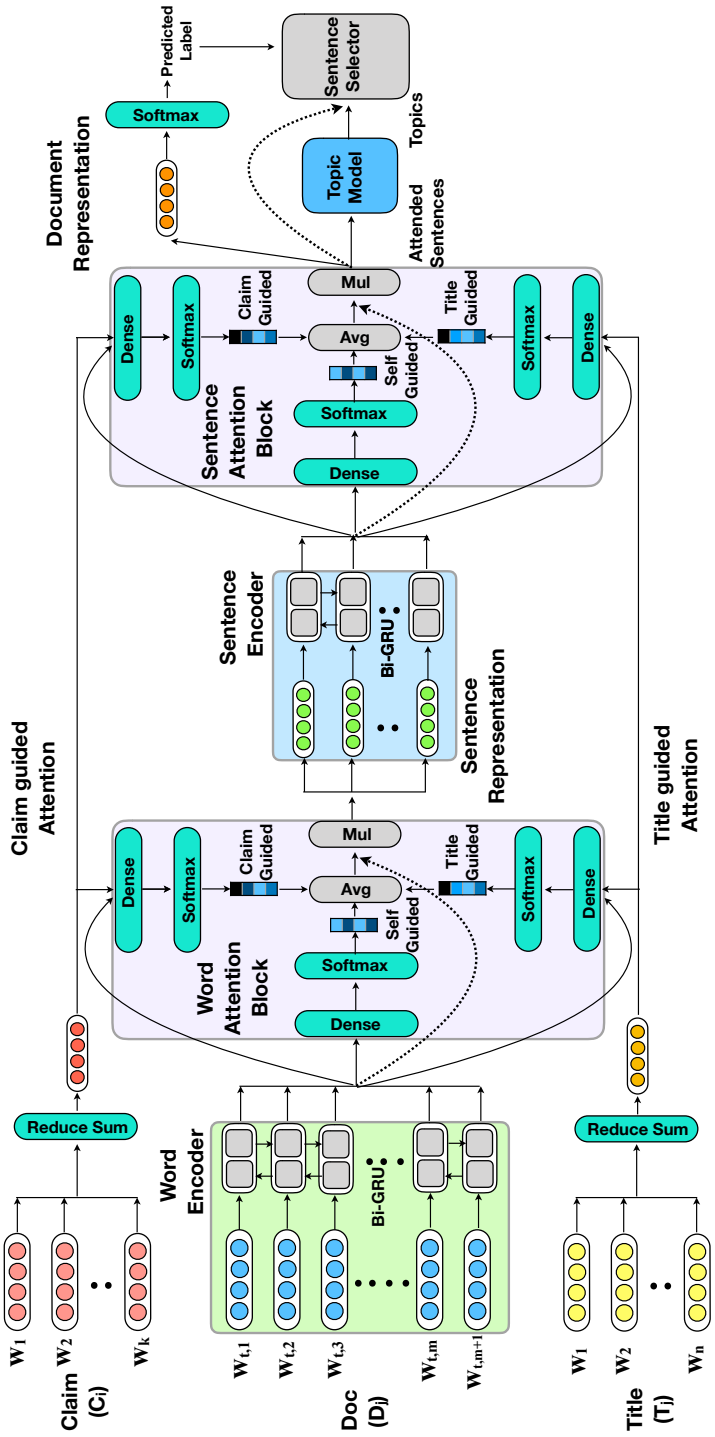



Figure 2: SUMO's neural network architecture for establishing the correctness of Web claims.

**Claim-driven Hierarchical Attention.**, aims to attend salient words that are significant and have relevance to the content of the claim. Similarly, we aim to attend the salient sentences at the sentence level attention. Recent works have used claim guided attention to model the contextual representation of the retrieved documents from the Web. These approaches provide claim-guided attention by first concatenating the claim word embeddings with document word embeddings and then applying a dense softmax layer to learn the attention weights as follows:

$$\begin{aligned} r_i = c_i \parallel d_i \;\; \& \;\; a_i = \tanh(W_a r_i + b_a) \\ \alpha = \text{softmax}(a_i), \end{aligned} \tag{5.1}$$

where $c_i$ and $d_i$ are the $i^{th}$ claim and document embeddings. $W_a$ and $b_a$ are the weight matrix and bias and $\alpha$ is the learned attention weight. However, during experiments, we observe that applying claim-based attention provides an inferior overall document representation. Therefore, we do not concatenate the claim and document embeddings before attention weight computation.

Each claim $c_i$ is consists of $l$ maximum number of words as $\{w_1, w_2, ....., w_l\}$. We represent each claim $c_i$ as the summation of embeddings of all the words contained in it as: $Cl_i = \sum_{j=1}^{l} f(w_j)$, where $f(w_j)$ is the word embedding of the $j_{th}$ word of claim $c_i$. Claim representation $Cl_i$ and hidden states $h_j$ from the GRU are used to **compute word-level claim-driven attention weights** as:

$$\begin{aligned} u_{j,i} = \tanh(W_{j,i} h_j + b_{j,i}) \\ \alpha^C_{j,i} = \text{softmax}(u^\top_{j,i} Cl_i), \end{aligned} \tag{5.2}$$

where $W_{j,i}$ and $b_{j,i}$ are the weight matrix and bias, $\alpha^C_{j,i}$ is the word level claim driven attention weight vector, and $h_j = (h_{j,1}, h_{j,2}, ..., h_{j,l})^\top$ represents the tuple of all the hidden states of the words contained in the $j^{th}$ sentence. To **compute sentence level claim-driven attention weights**, we use claim representation $Cl_i$ and hidden states $h^S_j$ from the sentence level GRU units as concatenations of both forward and backward hidden states $h^S_j = \overrightarrow{h^S_j} \parallel \overleftarrow{h^S_j}$ as follows:

$$\begin{aligned} u_j &= \tanh(W_j h^S + b_{j,i}) \\ \alpha_j^C &= \text{softmax}(u_j^\top C l_i), \end{aligned} \tag{5.3}$$

where $W_j$ and $b_j$ are the weight matrix and bias, $h^S = (h_1^S, h_2^S, ..., h_l^S)^\top$ is the combination of all hidden states from sentences, and $\alpha_j^C = (\alpha_{j,1}, \alpha_{j,2}, ..., \alpha_{j,k})^\top$ is the sentence level claim-driven attention weight vector for the $j^{th}$ document.

**Title-driven Hierarchical Attention.** The objective of using the document title is to guide the attention in capturing sections in the document that are more critical and relevant for the title. Articles convey multiple perspectives, often reflected in their titles. By title-driven attention, we attend to those words and sentences that are not covered in claim-driven attention. Title-driven attention at both word and sentence level can be computed in a similar fashion as claim-driven attention. Each title $t_i$ is comprised of $l$ maximum number of words as $\{w_1, w_2, \ldots, w_l\}$. We represent each claim $t_i$ as the summation of embeddings of all the words contained in it as: $T_i = \sum_{j=1}^{l} f(w_j)$. Title-driven attention weights for both words and sentence level can be computed as follows:

$$\begin{aligned} u_{j,i} &= \tanh(W_{j,i} h_j + b_{j,i}) \\ \alpha_{j,i}^T &= \text{softmax}(u_{j,i}^\top T_i) \\ u_j &= \tanh(W_j h^S + b_{j,i}) \\ \alpha_j^T &= \text{softmax}(u_j^\top T_i). \end{aligned} \tag{5.4}$$

**Hierarchical Self-Attention.** Self-attention is a simplistic form of attention. It tries to attend salient words in a sequence of words and salient sentences in a collection of sentences based on the self context of a sequence of words or a collection of sentences. In addition to claim-driven and title-driven attention, we apply self-attention to capture the unattended words and sentences which are not related to claim or title directly but are very useful for classification and summarization. Self-attention weights for both words and sentence level can be computed as follows:

$$
\begin{aligned}
u_{j,i} &= \tanh(W_{j,i}h_j + b_{j,i}) \\
\alpha_{j,i}^{Sl} &= \text{softmax}(u_{j,i}^{\top}) \\
u_j &= \tanh(W_j h^S + b_{j,i}) \\
\alpha_j^{Sl} &= \text{softmax}(u_j^{\top}),
\end{aligned}
\tag{5.5}
$$

where $\alpha_{j,i}^{Sl}$ and $\alpha_j^{Sl}$ are the self-attention weight vectors at word and sentence levels respectively.

**Fusion of Attention Weights.** We combine the attention weights from the three kinds of attention mechanisms: claim-driven, title-driven, and self-attention at both the word and sentence levels. At the word level, we set:

$$\alpha_j = (\alpha_{j,i}^{C} + \alpha_{j,i}^{T} + \alpha_{j,i}^{Sl})/3 \tag{5.6}$$

$$S_j = \alpha_j^{\top} h_j, \tag{5.7}$$

where $\alpha_{j,i}^{C}$, $\alpha_{j,i}^{T}$, and $\alpha_{j,i}^{Sl}$ are the attention weight vectors from claim, title and self-attention at the word level. $S_j$ is the formed sentence representation after overall attention for the $j^{th}$ sentence. At the sentence level, we set:

$$\alpha_j^{S} = (\alpha_j^{C} + \alpha_j^{T} + \alpha_j^{Sl})/3 \tag{5.8}$$

$$doc = \alpha_j^{\top} h^S, \tag{5.9}$$

where $\alpha_j^{C}$, $\alpha_j^{T}$, and $\alpha_j^{Sl}$ are the attention weight vectors from claim, title, and self-attention at the sentence level, and $doc$ is the formed document representation after overall attention.

**Attention on top of Attention (Atop).** Although the fusion of the three kinds of attention weights as an average of them works well, we realize that we lose some context by averaging. To deal with this issue, we use a novel attention on top of attention (Atop) method. We concatenate all three kinds of attentions $\alpha_{con}$ and $\alpha_{con}^{S}$ at both the word and sentence levels correspondingly. We apply a $\tanh$ activation based dense layer as a

scoring function and subsequently, a softmax layer to compute attention weights for each of three kinds of attention:

$$\begin{aligned}
\text{At word level:} \quad \alpha_{con} &= (\alpha^C_{j,i} \parallel \alpha^T_{j,i} \parallel \alpha^{Sl}_{j,i}) \\
u_{wa} &= \tanh(W_{wa}\alpha_{con} + b_{wa}) \\
\beta^w &= \text{softmax}(u_{wa}) \\
S_j &= \beta^w_1 \alpha^C_{j,i} + \beta^w_2 \alpha^T_{j,i} + \beta^w_3 \alpha^{Sl}_{j,i} \\
\text{At sentence level:} \quad \alpha^S_{con} &= (\alpha^C_j \parallel \alpha^T_j \parallel \alpha^{Sl}_j) \\
u_{sa} &= \tanh(W_{sa}\alpha^S_{con} + b_{sa}) \\
\beta^s &= \text{softmax}(u_{sa}) \\
doc &= \beta^s_1 \alpha^C_j + \beta^s_2 \alpha^T_j + \beta^s_3 \alpha^{Sl}_j,
\end{aligned} \tag{5.10}$$

where $\beta^w$ and $\beta^s$ are the learned attention weight vectors for three kinds of attentions at the word and sentence levels, and $doc$ is the formed document representation after Atop attention.

**Prediction and Optimization.** We use the overall document representation $doc$ in a softmax layer for the classification. To train the model, we use standard softmax cross-entropy with logits as a loss function, we compute $\hat{y}$, the predicted label as:

$$\hat{y} = \text{softmax}(W_{cl} doc + b_{cl}). \tag{5.11}$$

## 3.2 Generating Explainable Summary

Recent works retrieve documents from the Web as external evidence to support or refute the claims and thereafter extract snippets as explanations to model's decision [**sadhan9**, 7]. However, the extracted snippets from these methods are often redundant and lack topical diversity. The objective of our summarization algorithm is to provide ranked list of sentences that are: novel, non-redundant, and diverse across the topics identified from the text of the documents. In this section, we outline the method we utilize for achieving this objective.

**Multi-topic Sentence Model:** Each sentence in the document that is retrieved against the claim is modeled as a collection of topics: $s =$

$\langle a^{(1)}, a^{(2)}, \ldots a^{(k)} \rangle$. Let $\mathcal{A}$ be the set of topics $a_i \in \mathcal{A}$ across all candidate sentences from all the pseudo relevant set of documents $D$ for the claim.

**Objective**. We formulate the summarization task as a diversification objective. Given a set of relevant sentences $\mathcal{R}$ which are attended by Atop attention in SUMO while establishing the claim's correctness. We have to find the *smallest* subset of sentences $\mathcal{S} \subseteq \mathcal{R}$ such that *all* topics $a_i \in \mathcal{A}$ are covered. This is a variation of the Set Cover problem [26, 28, 29, 25, 33, 32, 31]. However, unlike IA-Select [26] we do not choose to utilize the Max Coverage variation of the Set Cover problem. Instead, we formulate it as Set Cover itself [28, 29]. That is, given a set of topics $\mathcal{A}$, find a minimal set of sentences $\mathcal{S} \subseteq \mathcal{R}$ that cover those topics [29]. Additionally, the inclusion of each sentence in the subset $\mathcal{S}$ has a *cost* associated with it, given by:

$$\begin{aligned} cost(s) &= (Score)^{-1} \\ Score &= (\lambda\theta_s + (1-\lambda)(W_{wa} + W_{sa})), \end{aligned} \tag{5.12}$$

where $\theta_s$ is the topic distribution score for sentence $s$ computed using a topic model (e.g., Latent Dirichlet Allocation [27]), $W_{wa} = \sum_{i=1}^{l} W_{wa}(i)$ is the average of attention weights of the words contained in sentence $s$, $W_{sa}$ is the attention weight of the sentence $s$, and $\lambda$ is a parameter to be tuned. We briefly describe our adaptation of the Greedy algorithm, which provides an approximate solution to the Set Cover problem, based on the discussion in [28, 29, 25, 33, 32, 31].

[t]

# 4 Evaluation

**Datasets.** We use two publicly available datasets, namely PolitiFact political claims dataset and Snopes political claims dataset [7] for evaluating SUMO's capability for fact checking. Dataset statistics for both the datasets are shown in Table 5.1. In the case of Politifact, claims have one of the following labels, namely: 'true', 'mostly true', 'half true', 'mostly false', 'false', and 'pants-on-fire,'. We convert 'true', 'mostly true', and 'half true'

**Algorithm 2** Adaption of the approximate Greedy algorithm for Set Cover problem from [28, 29, 25, 33, 32, 31] to our topical diversification problem setting. At each iteration, a sentence is chosen that covers the most number of topics reflected by topic distribution score and has the highest attention weights. As an output, we are assured a non-redundant, novel, and a diversified set of sentences.

**Input:** $\mathcal{A}$: Set of topics learned from the topic model for diversification.
$\mathcal{R}$: Set of sentences, attended by Atop. **Output:** $\mathcal{S} \subseteq \mathcal{R}$: Diversified set of sentences over $\mathcal{A}$

$\mathcal{S} \leftarrow \phi$ ; // $\mathcal{S}$ contains diversified sentences

$\mathcal{A}' \leftarrow \phi$ ; // $\mathcal{A}'$ contains topics covered by $\mathcal{S}$

**while** $\mathcal{A}' \neq \mathcal{A}$ **do**

/* identify the sentence that covers the most topics and is highly relevant for fact-checking */

$s^* \leftarrow \arg\min_{s \in \mathcal{R} \setminus \mathcal{S}} \frac{cost(s)}{|\mathcal{A} - \mathcal{T}'|}$ $\mathcal{A}' \leftarrow \mathcal{A}' \cup \{a_{s^*}\}$ ; // $a_{s^*}$ is the dominant topic of sentence $s^*$

$\mathcal{S} \leftarrow \mathcal{S} \cup s^*$

**end**

Table 5.1: Dataset Statistics

| PUBLIC DATASETS | | |
|---|---|---|
| STATISTICS | POLITIFACT | SNOPES |
| #CLAIMS | 3568 | 4341 |
| #DOCUMENTS | 29556 | 29242 |
| #DOMAINS | 3028 | 3267 |
| NEW DATASETS | | |
| STATISTICS | CLIMATE | HEALTH |
| #CLAIMS | 104 | 100 |
| #DOCUMENTS | 1050 | 978 |
| #DOMAINS | 97 | 83 |

labels to the 'true' and the rest of them to 'false' label. For the Snopes dataset, each claim has either 'true' or 'false' as a label.

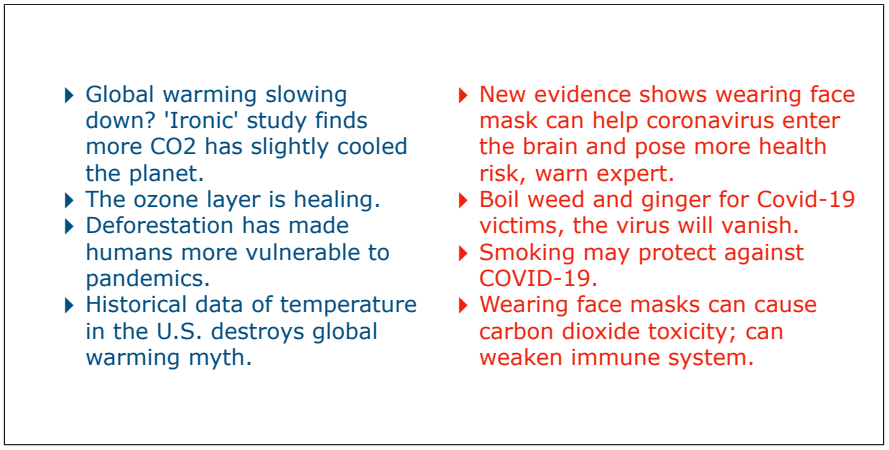



Figure 3: Examples from climate change and health care dataset

We evaluate SUMO for the task of summarization on PolitiFact, Snopes, Climate, and Health datasets. The two new datasets, Climate and Health, are about climate change and health care respectively. We test SUMO only on the PolitiFact and Snopes dataset for the task of fact checking as they are magnitudes larger than the new datasets that we release. The climate change dataset contains claims broadly related to climate change and global warming from `climatefeedback.org`. We use each claim as a query using Google API to search the Web and retrieve external evidences in the form of search results. Similarly, we create a dataset related to health care that additionally contains claims pertaining to the current global COVID-19 pandemic from `healthfeedback.org`. Examples of claims from these two datasets are shown in Figure 3. We make the new datasets, publicly available to the research community at the following URL: `https://github.com/rahulOmishra/SUMO/`.

**SUMO Implementation.** We use TensorFlow to implement [8] SUMO. We use per class accuracy and macro $F_1$ scores as performance metrics for evaluation. We use bi-directional Gated Recurrent Unit (GRU) with a hidden size of 200, word2vec [22], and GloVe [21] embeddings with embedding size of 200 and softmax cross-entropy with logits as the loss function. We keep the learning rate as 0.001, batch size as 64, and gradient

[8] `https://github.com/rahulOmishra/SUMO/`

clipping as 5. All the parameters are tuned using a grid search. We use 50 epochs for each model and apply early stopping if validation loss does not change for more than 5 epochs. We keep maximum sentence length as 45 and maximum number of sentences in a document as 35. For the task of summarization, we use Latent Dirichlet Allocation (LDA) [**blei**] as a topic model to compute topic distribution scores and the dominant topic for each candidate sentence.

# 5 Results

## 5.1 Setup for the Task of Claim Correctness

We experiment with five variants of our proposed SUMO model and compare with six state-of-the-art methods. The six state-of-the-art methods are as follows. First, we have the basic Long Short Term Memory (LSTM) [30]) unit which is used with claim and document contents for classification. Second, we have a convolutional neural network (CNN) [20] for document classification. Third, we compare against the model proposed in [19] that uses a hierarchical representation of the documents using hierarchical LSTM units (Hi-LSTM). Fourth, we compare against the model proposed in [17] that uses a hierarchical neural attention on top of hierarchical LSTMs (HAN) to learn better representations of documents for classification. Fifth, we compare against the model proposed in [7] that uses a claim guided attention method (DeClarE) for correctness prediction of claims in the presence of external evidences. Sixth and finally, we compare against the recent work [3] that improves on DeClarE method by using latent aspects (speaker, topic, or domain) based attention.

The proposed five variants of our method SUMO are as follows. First, we have the SUMO-AW2V variant that corresponds to the basic SUMO model with word2vec embeddings. Second, we have SUMO-AtopW2V variant consists of the SUMO model with WORD2VEC embeddings. Furthermore, in SUMO-AtopW2V we use Atop method of attention fusion rather than a simple average. Third, we have the SUMO-AGlove variant, which is the basic SUMO model that uses GloVe embeddings. Fourth, we have the SUMO-AtopGlove variant, that consists of the SUMO model

Table 5.2: Comparison of the proposed models with various state of the art baseline models for two publicly available datasets.

| POLITIFACT | | | |
|---|---|---|---|
| **Model** | **True Accuracy** | **False Accuracy** | **Macro $F_1$** |
| LSTM | 53.51 | 56.32 | 57.89 |
| CNN | 55.92 | 57.33 | 59.39 |
| HAN | 60.13 | 65.78 | 63.44 |
| DeClarE (full) | 68.18 | 66.01 | 67.10 |
| SADHAN-agg | 68.37 | 78.23 | 75.69 |
| SUMO-AW2V | 67.30 | 69.22 | 70.74 |
| SUMO-AtopW2V | 67.81 | 70.09 | 71.15 |
| SUMO-AGlove | 68.03 | 72.57 | 72.39 |
| SUMO-AtopGlove | 68.93 | 73.43 | 72.79 |
| SUMO-AtopGlove+source-Emb | **69.33** | **80.08** | **77.69** |
| SNOPES | | | |
| **Model** | **True Accuracy** | **False Accuracy** | **Macro $F_1$** |
| LSTM | 69.23 | 70.67 | 69.89 |
| CNN | 72.05 | 74.29 | 72.63 |
| HAN | 72.89 | 76.25 | 73.84 |
| DeClarE (full) | 60.16 | 80.78 | 70.47 |
| SADHAN-agg | 79.47 | 84.26 | 80.09 |
| SUMO-AW2V | 77.32 | 80.67 | 75.56 |
| SUMO-AtopW2V | 78.02 | 81.66 | 76.86 |
| SUMO-AGlove | 78.74 | 82.03 | 77.22 |
| SUMO-AtopGlove | 78.89 | 82.46 | 78.45 |
| SUMO-AtopGlove+source-Emb | **81.29** | **86.82** | **82.93** |

with GloVe embeddings. Moreover, in SUMO-AtopGlove, we use Atop method of attention fusion rather than a simple average. Fifth and finally, we have the SUMO-AtopGlove+source-Emb variant that is similar to SUMO-AtopGlove however with additional source embeddings (domains of retrieved documents).

## 5.2 Claim Correctness Task Results

The results for establishing claim correctness are shown in Table 2. We observe that the basic LSTM based model achieves $57.89\%$ and $69.89\%$ in terms of macro $F_1$ accuracy in prediction of claim correctness for POLITIFACT and SNOPES, respectively. The CNN model performs slightly better than LSTM as it captures the local contextual features better. The hierarchical attention network outperforms CNN with macro $F_1$ accuracy of $63.4\%$ and $73.84\%$. The reason for this improvement is hierarchical representation using word and sentence level attention. The state of the art DeClarE model provides significant improvements on baseline methods with macro $F_1$ accuracy of $67.10\%$ and $70.47\%$. This gain can be attributed to claim guided attention and source embeddings. However, we observe that this design of claim based attention is not very effective.

The more recent work, SADHAN improves on DeClarE, which uses a similar design for claim-oriented attention and incorporates a more comprehensive structure by using several latent aspects to guide attention. SADHAN outperforms DeClarE with macro $F_1$ accuracy of $75.69\%$ and $80.09\%$, respectively. Interestingly, we observe that the basic SUMO model with word2vec embeddings performs better than DeClarE with source embeddings. This observation is a clear indication of the superiority of our claim- and title-driven attention design. The SUMO with Atop attention fusion is more effective than a simple average fusion of attention weights, which becomes apparent from the gain in macro $F_1$ accuracy in both the datasets.

SUMO with pertained GloVe embeddings outperforms the word2vec versions of SUMO as the GloVe embeddings are trained on a large corpus and therefore captures better context for the words. SUMO-AtopGlove+sourceEmb outperforms all the other models and it is statistically significant with a p-value of $2.79 \times 10^{-3}$ for POLITIFACT and $3.09 \times 10^{-4}$ for SNOPES. The statistical significance values were computed using a two sample Student's t-test. We notice that SUMO could not outperform SADHAN without source embeddings, as SADHAN uses the very complex structure, having three parallel models with hierarchical latent aspects guide attention. However, SADHAN has many drawbacks. First, it is challenging to train and requires more hardware resources and time. Second, the

latent aspects are not available for all the Web claims. Therefore, it is not generalizable. Third, it fails to accommodate new values of latent variables at the test time.

## 5.3 Setup for the Task of Summarization

For the evaluation of the summarization capability of SUMO, we create gold reference summaries for claims. For creating the gold reference summaries, we include all the facts related to the claim, which are important for the claim correctness prediction, non-redundant, and topically diverse. We find that the descriptions provided for a claim on fact-checking websites such as `snopes.com` and `politifact.com` are suitable for this purpose. We use cosine similarity score of 0.4 between claims and sentences of description to filter out irrelevant or noisy sentences. As evaluation metrics, we use ROUGE-1, ROUGE-2, and ROUGE-L scores. The ROUGE-1 score represents the overlap of unigrams, while the ROUGE-2 score represents the overlap of bigrams between the summaries generated by the SUMO system and gold reference summaries. The ROUGE-L score measures the longest matching sequence of words using Longest Common Sub-sequence algorithm.

Standard summarization techniques are not useful in such a scenario as the objective of summarization with standard techniques is usually not fact-checking. Hence, we compare the SUMO results with an information retrieval (BM25) and a natural language processing based method (QuerySum). BM25 is a ranking function, which uses a probabilistic retrieval framework and ranks the documents based on their relevance to a given search query. We use Web claims as a query and apply BM25 to get the most relevant sentences from all the documents retrieved for the claim. We also compare the results with the query-driven attention based abstractive summarization method QuerySum [11], which also uses a diversity objective to create a diverse summary. We use ROUGE metrics with a gold reference summary to evaluate the generated summaries.

Table 5.3: Results for the Task of Summarization.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| BM25 | 26.08 | 14.78 | 29.98 |
| QuerySum | 29.78 | 16.49 | 30.16 |
| SUMO | **33.89** | **19.21** | **35.92** |

## 5.4 Comparison of Summarization Results

Results for the task of summarization are shown in Table 3, the QuerySum method performs significantly better than BM25 with a ROUGE-L score of 30.16 as it uses query-driven attention and diversity objective, which results in a diverse and query oriented summary. The proposed model SUMO outperforms QuerySum with a ROUGE-L score of 35.92. We attribute this gain to the use of word and sentence level weights, which are trained using back-propagation with correctness label. We also notice that in QuerySum some sentences are related to the claim but are not useful for fact checking. Therefore, they are absent in the gold reference summary. The results for SUMO are statistically significant ($p$-value $= 1.39 \times 10^{-4}$) using a pairwise Student's t-test.

# 6 Conclusion

We presented SUMO, a neural network based approach to generate explainable and topically diverse summaries for verifying Web claims. SUMO uses an improved version of hierarchical claim-driven attention along with title-driven and self-attention to learn an effective representation of the external evidences retrieved from the Web. Learning this effective representation in turn assists us in establishing the correctness of textual claims. Using the overall attention weights from the novel Atop attention method and topical distributions of the sentences, we generate extractive summaries for the claims. In addition to this, we release two important datasets pertaining to climate change and healthcare claims.

In future, we plan to investigate the BERT [2] and other Transformer [14] architecture based embedding methods in place of GloVe [21] embeddings

for better contextual representation of words.

# References

[1] **Pepa Atanasova, Jakob Grue, Simonsen Christina, and Lioma Isabelle**. "Generating Fact Checking Explanations." In: (2019). arXiv: arXiv:2004.05773v1.

[2] **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[3] **Rahul Mishra and Vinay Setty**. "SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection." In: ICTIR '19. Santa Clara, CA, USA, 2019, pp. 197–204. ISBN: 9781450368810.

[4] **Kai Shu, Suhang Wang, and Huan Liu**. "Beyond News Contents: The Role of Social Context for Fake News Detection." In: *WSDM*. ACM. 2019, pp. 312–320.

[5] **Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu**. "Unsupervised Fake News Detection on Social Media: A Generative Approach." In: AAAI '19. Feb. 2019.

[6] **Jing Ma, Wei Gao, and Kam-Fai Wong**. "Detect Rumor and Stance Jointly by Neural Multi-Task Learning." In: WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 585–593. ISBN: 9781450356404. DOI: 10.1145/3184558.3188729. URL: https://doi.org/10.1145/3184558.3188729.

[7] **Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum**. "DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning." In: *EMNLP*. 2018, pp. 22–32.

[8] **Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein**. "A Stylometric Inquiry into Hyperpartisan and Fake News." In: *ACL*. Vol. 1. 2018, pp. 231–240.

[9] **Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu**. "Neural User Response Generator: Fake News Detection with Collective User Intelligence." In: IJCAI '18. 2018, pp. 3834–3840.

[10] **Brian Xu, Mitra Mohtarami, and James Glass**. "Adversarial Domain Adaptation for Stance Detection." In: Nips (2018), pp. 1–6.

[11] **Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran**. "Diversity driven attention model for query-based abstractive summarization." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1063–1072. DOI: 10.18653/v1/P17-1098. URL: https://www.aclweb.org/anthology/P17-1098.

[12] **Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum**. "Where the truth lies: Explaining the credibility of emerging claims on the web and social media." In: *WWW*. 2017, pp. 1003–1012.

[13] **Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi**. "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. URL: https://www.aclweb.org/anthology/D17-1317.

[14] **Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin**. "Attention Is All You Need." In: abs/1706.03762 (2017).

[15] **Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha**. "Detecting Rumors from Microblogs with Recurrent Neural Networks." In: *IJCAI*. 2016, pp. 3818–3824.

[16] **Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit**. "A Decomposable Attention Model for Natural Language Inference." In: (2016). ISSN: 0001-0782. eprint: 1606.01933.

[17] **Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy**. "Hierarchical attention networks for document classification." In: *NAACL: HLT*. 2016, pp. 1480–1489.

[18] **Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong**. "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites." In: CIKM '15. Melbourne, Australia, 2015, pp. 1751–1754.

[19] **Duyu Tang, Bing Qin, and Ting Liu**. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1422–1432. DOI: 10.18653/v1/D15-1167.

[20] **Yoon Kim**. "Convolutional neural networks for sentence classification." In: *arXiv preprint arXiv:1408.5882* (2014).

[21] **Jeffrey Pennington, Richard Socher, and Christopher D. Manning**. "Glove: Global vectors for word representation." In: *In EMNLP*. 2014.

[22] **Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean**. "Distributed Representations of Words and Phrases and Their Compositionality." In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[23] **Carlos Castillo, Marcelo Mendoza, and Barbara Poblete**. "Information credibility on twitter." In: *WWW*. 2011.

[24] **Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei**. "Rumor Has It: Identifying Misinformation in Microblogs." In: EMNLP '11. 2011. ISBN: 9781937284114.

[25] **David P. Williamson and David B. Shmoys**. "The Design of Approximation Algorithms." In: New York, NY, USA: Cambridge University Press, 2011. ISBN: 0521195276, 9780521195270.

[26] **Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong**. "Diversifying Search Results." In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09. Barcelona, Spain: Association for Computing Machinery, 2009, pp. 5–14. ISBN: 9781605583907. DOI: 10.1145/1498759.1498766. URL: https://doi.org/10.1145/1498759.1498766.

[27] **David M. Blei, Andrew Y. Ng, and Michael I. Jordan**. "Latent Dirichlet Allocation." In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[28] **Bernhard Korte and Jens Vygen**. "Approximation Algorithms." In: *Combinatorial Optimization: Theory and Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 361–396. ISBN: 978-3-662-21711-5. DOI: 10.1007/978-3-662-21711-5_16.

[29] **Vijay V. Vazirani**. "Approximation Algorithms." In: New York, NY, USA: Springer-Verlag New York, Inc., 2001. ISBN: 3-540-65367-8.

[30] **Sepp Hochreiter and Jürgen Schmidhuber**. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

[31] **Vasek Chvátal**. "A Greedy Heuristic for the Set-Covering Problem." In: *Math. Oper. Res.* 4.3 (1979), pp. 233–235. DOI: 10.1287/moor.4.3.233. URL: https://doi.org/10.1287/moor.4.3.233.

[32] **László Lovász**. "On the ratio of optimal integral and fractional covers." In: *Discret. Math.* 13.4 (1975), pp. 383–390. DOI: 10.1016/0012-365X(75)90058-8. URL: https://doi.org/10.1016/0012-365X(75)90058-8.

[33] **David S. Johnson**. "Approximation Algorithms for Combinatorial Problems." In: *J. Comput. Syst. Sci.* 9.3 (1974), pp. 256–278. DOI: 10.1016/S0022-0000(74)80044-9. URL: https://doi.org/10.1016/S0022-0000(74)80044-9.

# Paper IV:
# MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching

# MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching

**Rahul Mishra**[1], **Piyush Yadav**[2], **Remi Calizzano**[3], **Markus Leippold**[4]

[1] Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway
[2] Lero, NUI Galway, Ireland
[3] DFKI, Germany
[4] University of Zurich, Switzerland

Published in the proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)

**Abstract:**
Measuring the congruence between two texts has several useful applications, such as detecting the prevalent deceptive and misleading news headlines on the web. Many works have proposed machine learning based solutions such as text similarity between the headline and body text to detect the incongruence. Text similarity based methods fail to perform well due to different inherent challenges such as relative length mismatch between the news headline and its body content and non-overlapping vocabulary. On the other hand, more recent works that use headline guided attention to learn a headline derived contextual representation of the news body also result in convoluting overall representation due to the news body's lengthiness. This paper proposes a method that uses inter-mutual attention-based semantic matching between the original and synthetically generated headlines, which utilizes the difference between all pairs of word embeddings of words involved. The paper also investigates two more variations of our method, which use concatenation and dot-products of word embeddings of the words of original and synthetic headlines. We observe that the proposed method outperforms prior arts significantly for two publicly available datasets.

# 1 Introduction

In the age of the prevalence of smart handheld devices, most of the information consumption is digital. This paradigm shift in the way people consume information has also brought forth several new challenges, such as misinformation and deceptive content. News headlines that incorrectly represent the content of the news body are called incongruent or click-baits. A deceptive and misleading news headline can result in false beliefs and wrong opinions. News titles play an essential role in making first impressions to readers and thereby deciding the viral potential of news stories within social networks [2]. Most users rely only on the news title content to determine which news items are significant enough to read [19]. The curse of deceptive content gets amplified by several magnitudes when people share it without reading news body content [19]. Consider


Figure 1: Examples of Incongruent Headlines related to politics and health-care.

an example of an incongruent headline in Figure 1 taken from CNN (cnn.com).[9] The headline states that “Trump says GOP working on tax plan for middle class" whereas body content mentions that “It’s unclear what tax proposal Trump was referring to on Saturday" which contradicts with the headline.

In another example in Figure 1,[10] the headline reads “The Scary New Science That Shows Milk Is Bad For You" but a part of the body text clearly states that “The study was small, to be sure, and it included no women." The headline radically generalizes the claim made in the study and exaggerates it.

Many machine learning based solutions have been investigated previously in the literature to detect click-baits and news headline incongruence. Some initial works [18][14] use sentence matching based methods to compute similarity or dissimilarity between the web claims and news headlines to detect incongruence. Researchers [21][12] have also utilized different classification methods and used linguistic and stylistic features to learn a classifier to identify the incongruity between news titles and news body. The authors in [5] propose neural attention-based methods to find the entailment between the news items’ title and body. Some recent works [7] have identified the significance of generative methods such as generative adversarial networks for incongruence detection.

Previous methods applied to click-bait detection that use natural language processing techniques are not suitable for the detection of headline incongruence as this problem requires more facets and aspects to be covered than just stylistic features [10]. The methods that use a text similarity based approach to detect incongruence perform poorly because of the long text content of the news body. Text similarity schemes work well in case of short texts.

This paper proposes a semantic matching technique based on inter-mutual attention that uses a synthetically generated headline corresponding to the news body content and original news headline to detect the incongruence.

[9] https://edition.cnn.com/2018/10/20/politics/donald-trump-tax-middle-income/index.html

[10] https://www.motherjones.com/environment/2015/11/dairy-industry-milk-federal-dietary-guidelines/

The proposed inter-mutual attention technique is inspired by some recent works, which use the intra-mutual word to word attention within a sentence to detect sarcasm [9] and intra-mutual user to user attention within a retweet propagation path sequence to detect misinformation [3].

In sum, the major contributions of this work are as follows:

(1) We are the first to use inter-mutual attention based semantic matching to detect incongruent news headlines. The key idea is to get the pairwise difference between word embeddings of original and synthetic headlines and compute a mutual attention matrix after applying a dense layer. Subsequently, row-wise max-pooling can be used to compute the attention scores, to be used for the classification.

(2) We use synthetic headlines generated from various generative adversarial networks based schemes using news body content to get the effective, contextual, and low dimensional representation.

(3) We also investigate two additional variants of the proposed model, which incorporate addition and concatenation of word embeddings of the word pairs of original and synthetic headlines.

(4) We combine all the three variants of word embedding operations in a clubbed model, which outperforms the three variants individually.

(5) We conduct experiments with two publicly available datasets, which show the effectiveness of the proposed models to detect incongruent news headlines.

# 2 Related Works

Most of the literature's initial works propose using linguistic and statistical features based classifiers to detect click-baits and incongruent headlines. The authors of [22] suggest to perform lexical and syntactic analysis for the identification of the click-baits. In [17], authors use linguistic features to learn support vector machine (SVM) based classifier to detect click-baits, and they also release a manually annotated dataset. Authors of [20] use text features and meta-information of tweets to learn a classifier to

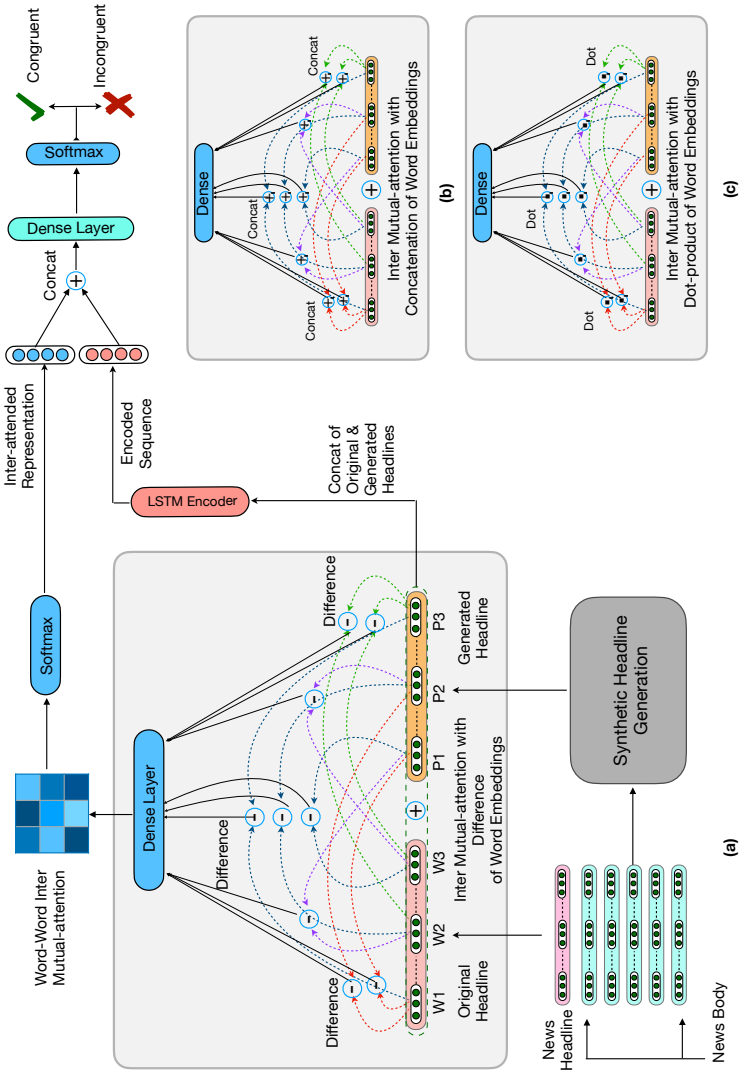



Figure 2: Overall Architecture of the MuSeM model. (a) A version of the MuSeM model with difference operation between word embedding pairs. (b) A depiction of concatenation operation between word embedding pairs. (c) A depiction of the dot-product operation between word embedding pairs.

detect click-baits.

On the other hand, some works [18][14] deal with sentence matching based stance classification which is a closely related problem to headline incongruence. These methods are not suitable directly for headline incongruence due to some challenges such as relative length mismatch between the news headline and its body content and non-overlapping vocabulary. Authors in [15] propose to use a co-training approach with myriad kinds of features such as, e.g., sentiments, textual, and informality. Some very recent works such as [5] use neural attention [1] based approach to achieve headline guided contextual representation of the news bod text. This headline guided attention also results in convoluting the body content's overall representation due to its lengthiness.

We propose an improved semantic matching between the news headline and its body text via inter-mutual attention. The inter-mutual attention based matching is performed between the original and a synthetically generated headline, rather than between original news headline and its body content. The synthetic headline is generated via state-of-the-art generative adversarial methods using the body contents of the news item. The generative adversarial techniques have been very successful in generating realistic artificial images [4], videos [13][6] and text [11] contents. One of the pioneer work in artificial text generation is [16], which solves the problem of generator differentiation by updating the gradient policy directly. The authors of [11] introduce a GAN based text generation method, which uses a third module, called MANAGER, which helps the generator network to utilize some leaked feature information from the discriminator network. In [7], authors introduce a synthetic headline generation method based on style transfer and generative adversarial network.

The key idea behind the inter-mutual attention method is inspired by two recent contributions [9][3]. There are some significant differences between these works and the proposed model. Firstly, they use intra-mutual word to word or user to user attention, which builds on word embedding pairs or user embedding pairs occurring within a sentence or retweet path sequence correspondingly to compute the attention scores. In contrast, the proposed model uses pairs of word embeddings of words that belong to two different sentences, i.e., original and synthetic headline.

Secondly, these previous methods use a concatenation of embedding vectors to form an overall pair embedding representation. In contrast, the proposed method uses the difference between the word embeddings to form an overall pair embedding representation. The intuition behind these differences is evident as we detect incongruence between two pieces of texts (headline and body), and computing the difference between word embedding pairs results in capturing the cues related to semantic similarity or dissimilarity between the two.

# 3 Problem Definition and Proposed Model

This section introduces the problem definition and presents the overall architecture of the proposed model MuSeM in detail. We describe the synthetic headline generation schemes, and then we provide details of the inter-attentive semantic matching technique. We also present two more variants of MuSeM model and devise a clubbed model, which combines all the three variants.

## 3.1 Problem Definition

Given a news item $n_i \in N$, having headline $h_i$ and body content $b_i$, we need to predict whether there is incongruence between $h_i$ and $b_i$ by classifying the news as either "Congruent(C)" or "Incongreunt(I)." The news headline $h_i$ consists of a sequence of $l$ words denoted as $h_i = \{w_{h1}, w_{h2}, ..., w_{hl}\} \in W$, and the news body content $b_i$ consists of a sequence of $m$ words denoted as $b_i = \{w_{b1}, w_{b2}, ..., w_{hm}\} \in W$.

## 3.2 Word Embedding Layer

For a news item $n_i$, the corresponding headline $h_i$ of length $l$ and body content $b_i$ of length $m$ are represented as $h_i = \{f(w_{h1}), f(w_{h2}), ..., f(w_{hl})\}$ where $\forall j, f(w_{hj}) \in \mathbb{R}^d$ is a word embedding vector of dimension $d$ for the $j^{th}$ word in headline $h_i$, and $b_i = \{f(w_{b1}), f(w_{b2}), ..., f(w_{hm})\}$ where $\forall k, f(w_{bk}) \in \mathbb{R}^d$ is a word embedding vector of dimension $d$ for the $k^{th}$ word in body content $b_i$. We experiment with pre-trained GLOVE embeddings for the evaluation.

## 3.3 Synthetic Headline Generation

As discussed in the introduction, the semantic similarity between the headline and news body content is the most significant incongruence indicator. Text similarity schemes work well in case of short texts. However, the major bottleneck to compute such a similarity measure is the news body's text length. For this reason, the methods which use a text similarity based approach to detect incongruence perform poorly. Generative adversarial networks (GANs) are recently getting much traction for various application scenarios, from the generation of realistic artificial images to compositions of meaningful poems. We utilize GANs based text generation techniques to generate a short synthetic headline for each of the news items using respective news body contents. The resulting synthetic headline is a low dimension contextual representation of the news body content.

### 3.3.1 Generative Adversarial Network

A typical generative adversarial network [24] comprises two neural nets, a discriminator and a generator. Both of these sub neural nets compete with each other. The generator network tries to maximize the classification error, while the discriminator network tries to minimize it. Both the generator and discriminator converge together and reach an equilibrium state.

### 3.3.2 Sequence Generative Adversarial Nets (SeqGAN)

SeqGAN[16] is a sequence generation method based on generative adversarial networks. The key difference between standard GANs and SeqGAN is that SeqGAN bypasses the generator differentiation problem associated with original GANs by directly performing a gradient policy update using a reinforcement learning (RL) based approach. In the $\theta$-parameterized generative model $G_\theta$ and start state $s_0$, the expected end reward for output $y_1$ is defined as:

$$J(\theta) = E[R_T|s_0, \theta] = \sum_{y_1 \in Y} G_\theta(y_1|s_0) \cdot Q_{D_\phi}^{G_\theta}(s_0, y_1), \tag{5.1}$$

where $Q_{D_\phi}^{G_\theta}(s_0, y_1)$ is an action-value function. A recurrent neural network (RNN) is used for the generative model for sequences, and a convolutional neural network (CNN) is used as the discriminator. We refer to the respective paper for more details.

### 3.3.3 Stylized Headline Generation (SHG)

Since incongruent headlines and click-baits usually follow a certain catchy and deceptive writing style, it can be advantageous to generate synthetic headlines that mimic these similar writing styles. Authors of very recent work SHG[7] propose a style transfer based headline generation approach, which uses a generative adversarial network with style discriminator. A gated recurrent unit (GRU) based recurrent neural network (RNN) is used as a generator that tries to minimize the following negative log-likelihood:

$$L_G(\theta_G) = E_{(x,h)\in S}[-\log_{PG}(h|y^L, z)]. \tag{5.2}$$

Three variants of discriminators are used, one for distinguishing the styles of the original and generated headline, a second one to maintain the aligned distributions in both the original and generated headlines, and a third for making sure that headline and body pairs are correctly classified. Again, we refer to the respective paper for more details.

## 3.4 Inter-mutual Attentive Semantic Matching

We now discuss the proposed semantic sentence matching scheme in detail. Our key idea is to model the relationships between all possible pairs of the words occurring in both the sentences, which captures the inherent semantic similarity between the sentences. We use pretrained word embeddings vectors and apply a novel inter mutual attention technique to model the similarity or dissimilarity relationship between pair of sentences.

### 3.4.1 Word to Word Inter Mutual Attention

We compute mutual attention scores between a pair of sentences, of which first $h_i^o = \{f(w_{h^o1}), f(w_{h^o2}), ..., f(w_{h^ol})\}$ where $\forall j, f(w_{h^oj}) \in \mathbb{R}^d$ is the

original headline of the news item, represented in terms of a sequence of word embeddings, and second $h_i^s = \{f(w_{h^s1}), f(w_{h^s2}), ..., f(w_{h^sp})\}$ where $\forall j, f(w_{h^sj}) \in \mathbb{R}^d$ is the synthetically generated headline for news item $n_i$, represented in terms of a sequence of word embeddings. Here, $l$ and $p$ are lengths of the original headline $h_i^o$ and the synthetic headline $h_i^s$, respectively.

First of all, we compute the difference between word embedding vectors of each candidate word pair $W_q, W_r$. Candidate pairs are formed by selecting all possible combinations of the inter sentence word pairs, such as $W_q, W_r$, where $W_q$ and $W_r$ are the $q^{th}$ word of the original headline $h_i^o$ and the $r^{th}$ word of the synthetic headline $h_i^s$, respectively. Now we use a dense layer to project the difference of candidate embedding pairs into a scalar score:

$$C_{qr} = \theta_{diff}([f(w_{h^oq}) - f(w_{h^sr})]) + b_{diff}, \tag{5.3}$$

where $\theta_{diff} \in \mathbb{R}^{d\times 1}$ is a weight matrix and $b_{diff} \in \mathbb{R}$ a bias term. The score matrix $C = (C_{qr})$ is of dimension $l \times p$. We use row-wise avg-pooling and apply softmax to compute inter mutual attention scores for the original headline:

$$A^o = \underset{row}{\text{Softmax}}(\text{avg}\, C), \tag{5.4}$$

where $A_o$ is the learned inter-mutual attention weight vector for original headline. We use column-wise avg-pooling and apply softmax to compute inter mutual attention scores for synthetic headline.

$$A^s = \underset{col}{\text{Softmax}}(\text{avg}\, C), \tag{5.5}$$

where $A^s$ is the learned inter-mutual attention weight vector for synthetic headline. Subsequently, inter-mutual attended representations for original headline $M_{A^o}$ and for synthetic headline $M_{A^s}$ can be computed as:

$$
\begin{aligned}
M_{A^o} &= \sum_{i=1}^{l} f(w_{h^o i}) A^o{}_i \\
M_{A^s} &= \sum_{i=1}^{p} f(W_{h^s i}) A^s{}_i,
\end{aligned}
\tag{5.6}
$$

where $f(w_{h^o i})$ and $f(W_{h^s i})$ are the word embeddings of the $i^{th}$ word of the original and synthetic headline, respectively. Now, we compute the overall inter-mutual attended representation $M_A$ as:

$$
M_A = M_{A^o} + M_{A^s} \tag{5.7}
$$

### 3.4.2 Variants of Inter Mutual Attention

Although the MuSeM model with difference of word embeddings of candidate word pair works well, we also investigate and experiment with two other versions of MuSeM that use the dot product of candidate word embedding pairs and concatenation of word embedding pairs, similar to [8]. We notice that the dot-product and concatenation variants do not perform better than the difference-oriented variant in our experiments. We also combine all three operations, namely difference, dot-product, and concatenation, which performs slightly better than the purely difference oriented model. Attention scores for the dot-product and concatenation oriented models can be computed in a very similar fashion to the difference model, except for equation (5.3), which needs to be replaced by:

$$
\begin{aligned}
C_{qr} &= \theta_{dot}([f(w_{h^o q}) \cdot f(w_{h^s r})]) + b_{dot} \\
C_{qr} &= \theta_{con}([f(w_{h^o q}) \parallel f(w_{h^s r})]) + b_{con}.
\end{aligned}
\tag{5.8}
$$

For the sake of brevity, we do not repeat all the other equations.

### 3.4.3 Clubbed Model

We also experiment with a combined model in which all three variants namely difference based, dot-product based and concatenation based

methods are used in parallel and resulted overall embedding representation is used for computing attention scores.

$$\begin{aligned} F_{qr}^{dot} &= [f(w_{h^o q}) \cdot f(w_{h^s r})] \\ F_{qr}^{con} &= [f(w_{h^o q}) \parallel f(w_{h^s r})] \\ F_{qr}^{diff} &= [f(w_{h^o q}) - f(w_{h^s r})] \\ C_{qr} &= \theta_{dpc}([F_{qr}^{dot} \parallel F_{qr}^{con} \parallel F_{qr}^{diff}]) + b_{dpc}, \end{aligned} \tag{5.9}$$

where $C_{qr}$ is overall attention score, $\theta_{dpc}$ is the weight matrix and $b_{dpc}$ is the bias term for the clubbed model.

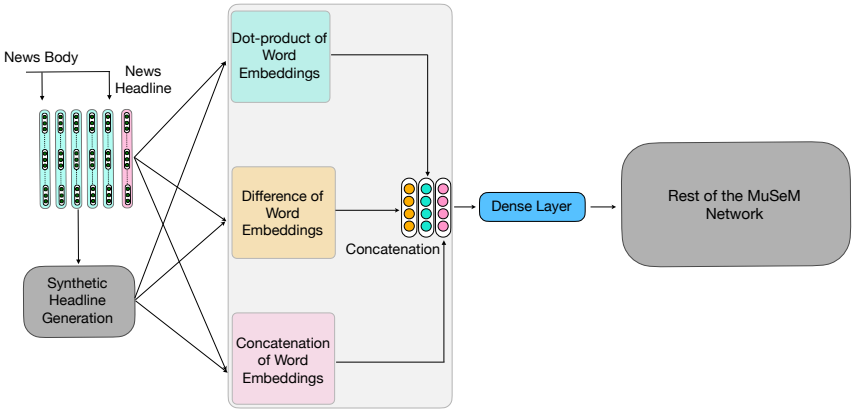



Figure 3: Clubbed Model Architecture

## 3.5 LSTM based Sequence Encoder

We concatenate the sequence of word embeddings of the words of the original headlines and words of the synthetically generated headline and use long short-term memory unit (LSTM) [26] to encode this overall sequence by using the standard LSTM equations as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i.[h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C.[h_{t-1}, x_t] + b_C) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o) \\
h_t &= o_t * \tanh(C_t),
\end{aligned}
\tag{5.10}
$$

where $h_{t-1}$ is previous hidden state and $x_t$ is the current input. We use the last hidden state of LSTM as the encoded representation of the overall sequence $M_E$.

### 3.6 Classification

We learn a joint representation $M$ of inter-mutual attended representation $M_A$ and encoded representation of the overall sequence $M_E$ using a nonlinear transformation layer with ReLU activation. Subsequently, we use a softmax layer to predict the label:

$$
\begin{aligned}
M &= \mathrm{ReLU}(W_t([M_A, M_E] + b_t) \\
\hat{y} &= \mathrm{Softmax}(W_{cl}M + b_{cl}),
\end{aligned}
\tag{5.11}
$$

where $W_{cl}$, $b_{cl}$, and $\hat{y}$ are the weight matrix, the bias term, and the predicted label, respectively. We use softmax cross-entropy with logits as loss $L$:

$$
L = -\sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}}. \tag{5.12}
$$

## 4 Experimental Setup

We implement [11] the proposed models within the TensorFlow framework. For the evaluation and comparison, we use Macro F1 and AUC scores as

[11] https://github.com/rahulOmishra/MuSem

evaluation metrics. All the parameters are tuned using a grid search. As a result, we keep learning rate as 0.001, batch size as 100, hidden states of LSTM as 100, and pretrained GloVe[25] embeddings of 300 dimensions. We use softmax cross-entropy with logits as the loss function, maximum sentence length as 50, dropout rate as 0.2, and the number of epochs as 10. For the synthetic headline generation, we use the default parameter values used in the papers mentioned above.

## 4.1 Dataset Statistics

We use two publicly available datasets, NELA17[12] and Click-bait Challenge[13] for the evaluation and comparison of the proposed model with the baseline methods. The NELA17 dataset is provided by [5]. Although they do not evaluate their model with it, they provide a script[14] to generate the dataset from an original news collection dataset. It contains 91042 news items in total, of which 45521 news items are congruent, and 45521 news items are non-congruent. The Click-bait Challenge dataset is a collection of social media posts, which are annotated as click-bait or non-click-bait using a crowd-sourcing platform via majority voting. It contains 21033 social media posts in total, of which 16150 posts are congruent, and 4883 posts are non-congruent.

Table 5.1: Dataset Statistics

| **Statistics** | **NELA17** |
|---|---|
| Total | 91042 |
| Non-congruent | 45521 |
| Congruent | 45521 |

The NELA17 dataset is balanced but the Click-bait Challenge dataset has class imbalance problem. We use the class_weight[15] parameter to adjust

[12] `https://github.com/BenjaminDHorne/NELA2017-Dataset-v1`

[13] `http://www.clickbait-challenge.org/`

[14] `https://github.com/sugoiii/detecting-incongruity-dataset-gen`

[15] `https://www.tensorflow.org/tutorials/structured_data/imbalanced_data`.

| Statistics | Click-bait Challenge |
|---|---|
| Total | 21033 |
| Non-congruent | 4883 |
| Congruent | 16150 |

the imbalance in the dataset.

# 5 Evaluation and Discussion

In this section, we compare the performance of the proposed models with state of the art methods and baseline models on two publicly available datasets. Subsequently, we also provide some explanations on the differences in the performance of the analyzed models. We also discuss a potentially useful trick from [3], which can improve the incongruence detection accuracy by capturing the newly uncovered cues.

We compare the proposed model with these baselines:

- **SVM:**[27] In this method, we use handcrafted linguistic and other statistical features to learn a classifier with support vector machine technique.
- **LSTM:**[26] In this method, we use long short term memory unit to encode both headline and body pair and apply softmax for the classification.
- **Hi-LSTM:**[23] This technique uses an LSTM based hierarchical encoder, which first encodes words and then uses another LSTM encoder to form the sentence representations.
- **Yoon:**[5] This is a state of the art, hierarchical dual encoder based model which uses headline guided attention to learn the contextual representation.

We experiment with four variants of the proposed MuSeM model.

- **MuSeM_diff_SeqGAN:** This is the MuSeM model with the difference between word embeddings, and synthetic headlines are generated using SeqGAN method.

- **MuSeM_dpc_SeqGAN:** This is the MuSeM model which uses a combination of all three operators, namely difference, concatenation, and dot-product between word embeddings and synthetic headlines are generated using SeqGAN method.
- **MuSeM_diff_SHG:**This is the MuSeM model with the difference between word embeddings, and synthetic headlines are generated using SHG method.
- **MuSeM_dpc_SHG:** This is the MuSeM model which uses a combination of all three operators, namely difference, concatenation, and dot-product between word embeddings and synthetic headlines are generated using SHG method.

Table 5.2: Comparison of the proposed models with various state of the art baseline models for the NELA17 Dataset.

| NELA17 Dataset | | |
|---|---|---|
| **Model** | **Macro F1** | **AUC.** |
| SVM | 0.622 | 0.637 |
| LSTM | 0.642 | 0.663 |
| HiLSTM | 0.651 | 0.672 |
| Yoon | 0.685 | 0.697 |
| MuSeM_diff_SeqGAN | 0.713 | 0.720 |
| MuSeM_dpc_SeqGAN | 0.719 | 0.727 |
| MuSeM_diff_SHG | 0.740 | 0.753 |
| MuSeM_dpc_SHG | **0.752** | **0.769** |

## 5.1 Results for NELA17 Dataset

In the case of the NELA17 dataset, the SVM model with linguistic and statistical features achieves $0.622$ and $0.637$ in terms of Macro F1 and AUC, respectively. The LSTM model outperforms SVM with Macro F1 as $0.642$ and AUC as $0.663$. This gain can be attributed to the suitability of LSTM to learn the contextual representation of text sequences. The Hierarchical LSTM (Hi-LSTM) performs slightly better than simple LSTM with Macro F1 as $0.651$ and AUC as $0.672$. The probable reason for this

Table 5.3: Comparison of the proposed models with various state of the art baseline models for the Click-bait challenge 2017 Dataset.

| Click-bait challenge 2017 Dataset | | |
|---|---|---|
| **Model** | **Macro F1** | **AUC.** |
| SVM | 0.618 | 0.629 |
| LSTM | 0.630 | 0.641 |
| HiLSTM | 0.642 | 0.656 |
| Yoon | 0.660 | 0.678 |
| MuSeM_diff_SeqGAN | 0.677 | 0.683 |
| MuSeM_dpc_SeqGAN | 0.690 | 0.698 |
| MuSeM_diff_SHG | 0.729 | 0.734 |
| MuSeM_dpc_SHG | **0.735** | **0.747** |

improvement is the better representation learned by H-LSTM, in the form of the documents' hierarchical structure. Since the Yoon model uses a dual hierarchical encoder, which encodes words and paragraphs of the new body text separately using an attention mechanism guided by news headline, it outperforms the Hi-LSTM with significant gains. The Yoon[16] model works well for long texts as it selects important paragraphs from the long body text, which reduces the effective size of the document. On the other hand, SVM, LSTM, and Hi-LSTM models do not scale well for long text sequences.

We compare the baselines and state-of-the-art models with four variants of the proposed model MuSeM. The MuSeM_diff_SeqGAN model performs better than the Yoon model with $0.713$ and $0.720$ in terms of Macro F1 and AUC. This performance improvement can be credited to low dimension representation of the news body content in the form of synthetic headline and inter-mutual attention based semantic matching. Although the headline guided attention to select relevant paragraphs in the Yoon model reduces the effective document length, the resultant document representation is still not of very low dimension. In contrast, the MuSeM model uses a very low dimensional representation of news body content in the form of a synthetic headline, which is more effective in semantic

[16]`https://github.com/david-yoon/detecting-incongruity`

matching.

MuSeM captures both compositional aspects and latent cues related to similarity and dissimilarity relationships among the words using LSTM based sequence encoder and inter-mutual attention based semantic matching. The MuSeM_dpc_SeqGAN model performs slightly better than MuSeM_diff_SeqGAN due to additional patterns pertaining to similarity relationship among words, captured by the combination of all three variants, i.e., difference based, dot-product based, and concatenation based methods. The MuSeM_diff_SHG model, which uses a style transfer based headline generation method, outperforms both variants of the MuSeM models with SeqGAN by a large margin. This performance improvement can be attributed to a better headline generation by stylized headline generation (SHG) method. The MuSeM_dpc_SHG model outperforms all the other models by achieving $0.752$ and $0.769$ in Macro F1 and AUC, respectively.

## 5.2 Results for Click-bait Challenge Dataset

In the case of the Click-bait Challenge dataset, we observe very similar trends as with the NELA17 dataset. The deep learning based methods perform significantly better than the handcrafted feature based SVM model. With $0.660$ and $0.678$ in Macro F1 and AUC, the Yoon model outperforms both the LSTM and Hi-LSTM models. All variants of the MuSeM model perform better than all the baseline methods, whereas MuSeM_dpc_SHG model achieves $0.735$ and $0.747$ in terms of Macro F1 and AUC, beating all the other variants.

## 5.3 Higher Order Inter-mutual Attention

Authors in [3] propose a variant of mutual attention which can be used to model higher-order relationship among the candidate words. Essentially, the proposed inter-mutual attention only models the relationship between two words at a time individually, and the presence of other words in the sentences are not taken into account. If we can also reckon the presence of other words during the computation of the inter-mutual attention, we can capture the additional contextual cues related to similarity or dissimilarity

relationship. The higher-order mutual attention trick allows us to model such relationships.

There are certain challenges to be taken care of before the application of higher order mutual attention to incongruence detection problem. Firstly, the original higher order mutual attention is designed for a single sequence as it uses intra-mutual attention within a sequence, while headline incongruence detection involves two sentences or sequences. Secondly, higher order mutual attention is achieved by applying a square or cube of the attention score matrix, which is possible only if the original score matrix is a square matrix. In the headline incongruence detection task, the score matrix may or may not be a square matrix because original and synthetically generated headlines may have different lengths.

# 6 Conclusion

This paper proposes inter-mutual attention based semantic matching (MuSeM) to detect incongruence in news headlines. We also investigate a different variant of MuSem, which combines three operations on word embedding pairs to compute inter-mutual attention scores. The proposed models outperform all the baselines in experiments with two publicly available datasets. We notice that the performance of inter-mutual attention based semantic matching greatly depends on the accuracy of synthetic headline generation step.

In future research, we plan to use the higher order word to word attention trick used in [3] to model and capture the indirect relationships between the word pairs. We plan to investigate and devise an end to end version of MuSeM model, where synthetic headline generation and semantic matching steps are seamlessly integrated. We also plan to conduct an analysis of the efficacy of the attention mechanism by visualizing the attention weights.

# References

[1] **Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio**. *Neural Machine Translation by Jointly Learning to Align and Translate*. In ICLR 2015 as oral presentation.

[2] **Julio Reis, Pedro Olmo, Raquel Prates, Haewoon Kwak, and Jisun An**. "Breaking the News : First Impressions Matter on Online News." In: (), pp. 357–366.

[3] **Rahul Mishra**. "Fake News Detection Using Higher-Order User to User Mutual-Attention Progression in Propagation Paths." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[4] **T. Karras, S. Laine, and T. Aila**. "A Style-Based Generator Architecture for Generative Adversarial Networks." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4396–4405.

[5] **Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung**. "Detecting Incongruity between News Headline and Body Text via a Deep Hierarchical Encoder." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 791–800.

[6] **Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li**. "PaGAN: Real-Time Avatars Using Dynamic Textures." In: *ACM Trans. Graph.* 37.6 (Dec. 2018). ISSN: 0730-0301. DOI: 10.1145/3272127.3275075. URL: https://doi.org/10.1145/3272127.3275075.

[7] **K. Shu, S. Wang, T. Le, D. Lee, and H. Liu**. "Deep Headline Generation for Clickbait Detection." In: *2018 IEEE International Conference on Data Mining (ICDM)*. 2018, pp. 467–476.

[8] **Yi Tay, Luu Anh Tuan, and Siu Cheung Hui**. "Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference." In: *Proceedings of EMNLP 2018*. 2018.

[9] **Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su**. "Reasoning with sarcasm by reading in-between." In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2018), pp. 1010–1020.

[10] **Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver**. "Incongruent Headlines: Yet Another Way to Mislead Your Readers." In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 56–61. DOI: 10.18653/v1/W17-4210. URL: https://www.aclweb.org/anthology/W17-4210.

[11] **Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang**. "Long Text Generation via Adversarial Training with Leaked Information." In: AAAI (2017).

[12] **Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf**. "BaitBuster : A Clickbait Identification Framework." In: (2017), pp. 8216–8217. arXiv: arXiv:1607.04606.

[13] **Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman**. "Synthesizing Obama: Learning Lip Sync from Audio." In: *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073640. URL: https://doi.org/10.1145/3072959.3073640.

[14] **Zhiguo Wang, Wael Hamza, and Radu Florian**. "Bilateral Multi-Perspective Matching for Natural Language Sentences." In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 4144–4150. DOI: 10.24963/ijcai.2017/579. URL: https://doi.org/10.24963/ijcai.2017/579.

[15] **Wei Wei and Xiaojun Wan**. "Learning to Identify Ambiguous and Misleading News Headlines." In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 4172–4178. ISBN: 9780999241103.

[16] **Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu**. "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 2852–2858.

[17] **A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly**. "Stop Clickbait: Detecting and preventing clickbaits in online news media." In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016, pp. 9–16.

[18] **William Ferreira and Andreas Vlachos**. "Emergent: a novel dataset for stance classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1163–1168. DOI: 10.18653/v1/N16-1138. URL: https://www.aclweb.org/anthology/N16-1138.

[19] **Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout**. "Social Clicks: What and Who Gets Read on Twitter?" In: *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. SIGMETRICS '16. Antibes Juan-les-Pins, France: Association for Computing Machinery, 2016, pp. 179–192. ISBN: 9781450342667. DOI: 10.1145/2896377.2901462.

[20] **Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen**. "Clickbait Detection." In: *ECIR'16*. ECIR'16 1 (2016).

[21] **Jonas Nygaard Blom and Kenneth Reinecke Hansen**. "Click bait: Forward-reference as lure in online news headlines." English. In: *Journal of Pragmatics* 76 (Jan. 2015), pp. 87–100. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2014.11.010.

[22] **Yimin Chen, Niall J. Conroy, and Victoria L. Rubin**. "Misleading Online Content: Recognizing Clickbait as "False News"." In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. WMDD '15. Seattle, Washington, USA: Associa-

tion for Computing Machinery, 2015, pp. 15–19. DOI: 10.1145/2823465.2823467.

[23] **Duyu Tang, Bing Qin, and Ting Liu**. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1422–1432. DOI: 10.18653/v1/D15-1167.

[24] **Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio**. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680.

[25] **Jeffrey Pennington, Richard Socher, and Christopher D. Manning**. "Glove: Global vectors for word representation." In: *In EMNLP*. 2014.

[26] **Sepp Hochreiter and Jürgen Schmidhuber**. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

[27] **Corinna Cortes and Vladimir Vapnik**. "Support-Vector Networks." In: *Mach. Learn.* 20.3 (Sept. 1995), pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411.

# Paper V:
# POSHAN: Cardinal POS Pattern Guided Attention for News Headline Incongruence

# POSHAN: Cardinal POS Pattern Guided Attention for News Headline Incongruence

**Rahul Mishra**[1], **Shuo Zhang**[2]

[1] Department of Electrical Engineering and Computer Science,
University of Stavanger, Stavanger, Norway
[2] Bloomberg, United Kingdom

The paper is currently under review.

**Abstract:**
Automatic detection of click-baits and incongruent news headlines is crucial to maintain the reliability of the Web and has raised much research attention. However, most existing methods perform poorly when news headline contains contextually important cardinal values such as a quantity or an amount. In this work, we focus on this particular case and propose a neural attention based solution, which uses a novel cardinal **Part of Speech (POS)** tags pattern based **h**ierarchical **a**ttention **n**etwork, namely ***POSHAN***, to learn effective representations of sentences in the news article. In addition, we investigate a novel cardinal phrase guided attention, which uses word embeddings of the contextually important cardinal value and neighbouring words. In the experiments conducted on two publicly available datasets, we observe that the proposed method gives appropriate significance to cardinal values and outperforms all the baselines. An ablation study of the *POSHAN*, shows that the cardinal POS-tag pattern based hierarchical attention is very effective for the cases in which headline contains cardinal values.

| |
|---|
| ***Headline:*** *Immigration Expert : US Will Have 100 Million New Immigrants in Next 50 Years.* |
| ***Body:*** *These projections show that new immigrants and their descendants will drive most U.S. population growth in the coming 50 years, as they have for the past half-century. Among the projected 441 million Americans in 2065, 78 million will be immigrants and 81 million will be people born in the U.S. to immigrant parents.* |

Figure 1: Example of an Incongruent Headline. The headline says, in the next 50 years, there will be 100 million new immigrants but the news body quotes about only 78 million new immigrants.

# 1 Introduction

News titles expose the first impression to readers and decide the viral potential of news stories within social networks [2]. Most of the users only rely on the news title to decide what to read further [19]. A deceptive and misleading news title can lead to false beliefs and wrong opinions. It becomes inversely worse when users share the news on social media without reading the news body but only skimming through the news title. The news headlines, which are ambiguous, misleading, and deliberately made catchy to lure the users to click, are called incongruent headlines or click baits [15]. Figure 1 illustrates an example.

There is a line of study that has been investigated and analyzed in the literature [22, 17, 21, 18, 14, 15, 9, 7], witnessed by different techniques such as linguistic feature based methods [22], generative adversarial networks[9, 3], and hierarchical neural attention networks Yoon. For example, Yoon et. al.[7] propose a headline text guided neural attention network to compute an incongruence score between the news headline and the corresponding body text. They introduce a hierarchical attention based encoder, which encodes words of news body text at the word level to form paragraph representations and encodes paragraphs to form document representation. However, we observe that these prior works fail to generalize and perform adequately in cases where news headlines contain a significant numerical value. The numerical values can be in the

form of a currency amount, counts of people, months, years or objects, etc. For instance, in Figure 1 an excerpt from a news item is shown, in which the news headline *"Immigration Expert : US Will Have 100 Million New Immigrants in Next 50 Years."*, contains two contextually important numerical figures i.e. *"100 Million"*,*"50 Years."*. The headline mentions, there will be 100 million new immigrants, but the news body quotes only 78 million new immigrants. The headline is deliberately made contradictory and exaggerating to look more sensational. It's apparent from this example that numerical and cardinal values are useful and crucial cues of the congruence of the news headlines.

All of the prior works suffer from not giving enough importance to numerical values. The headline guided attention-based methods such as Yoon, fail to attend relevant words related to cardinal phrases as they do not treat them specifically. On the other hand, generative adversarial network-based methods, which generate a synthetic headline from news body text to augment the dataset or to use them for similarity matching with original headlines, also miss cardinal aspects in the synthetically generated headlines. Clearly, the news headlines having numbers are not trivial cases for incongruence detection, and in this paper, we try to deal with news headline incongruence detection with special focus to such cases.

The objective of this work is to devise an incongruence detection method, which not only performs better than previously proposed techniques but also resolves the deterioration of classification accuracy with the news items in which the headline contains cardinal values. In specific, we leverage a novel Cardinal Part-of-Speech Tag patterns to drive the hierarchical neural attention to capture salient and contextually important words and sentences at the word and sentence levels correspondingly. The key idea of using cardinal pos patterns such as $(NN : CD : JJ)$ or $(VBD : CD : CD)$ is to use them as latent features associated with news headlines and learn the contextual embeddings based on data samples containing the same cardinal pos patterns. The embeddings are used at the test time to drive the attention and capture the salient words and sentences, which are significant for cardinal values. In addition, we investigate a cardinal phrase guided attention mechanism and combine both with standard headline guided attention. To utilize the better contextual representation of words, we fine tune the pre-trained BERT model and extract the word

embeddings, which are fed to a Bi-LSTM based sequence encoder.

We conduct experiments with the subset of two publicly available datasets and achieve state-of-the-art performance. The proposed model *POSHAN* not only outperforms all the other methods in original datasets but also its performance does not deteriorate much compared to other models, with derived datasets, containing only those data samples, which have numerical values in the news headlines. We visualize the Cardinal POS Pattern embeddings and overall attention weights to further analyse the effectiveness of the proposed model. We observe that the Cardinal POS Patterns have formed clearly separated clusters in embedding space, which connote the congruence and incongruence labels. It is apparent from the visualization of overall attention weights, that *POSHAN* model successfully attends the contextually important cardinal phrases in addition to other significant words. In nutshell, the major contributions of this work are:

- We focus on the news headline incongruence detection when news headlines containing numbers, and propose the cardinal POS pattern guided attention (Section 3.4) baseline.
- We propose a cardinal phrase guided attention (Section 3.6) mechanism and combine the both cardinal POS pattern and cardinal phrase attention with standard headline guided attention (Section 3.7) in a joint model (Section 3.8).
- We incorporate the proposed hierarchical attention methods on top of a Bi-LSTM based sequence encoder (Section 3.3) which encodes the sequence of fine-tuned (Section 5.2) pre-trained BERT embeddings (Section 3.2) of the words.
- In the evaluation with two publicly available datasets (Table 5.4 and 5.5 ), the proposed techniques outperform the baselines and state-of-the-art methods.
- We visualize the Cardinal POS Pattern embeddings and overall attention weights and conduct error analysis to analyze the effectiveness of the proposed model, and verify the effectiveness.

# 2 Related Work

Detection and prevention of misinformation and deceptive content online has gained lots of traction recently. Incongruent news and click-baits are very common forms of deception and misinformation. Naturally, most of the prior works in this area have treated the click-baits or news incongruence detection task as a standard text classification problem. Majority of the initial works are feature engineering heavy [16], exploiting diverse features such as linguistic features, lexicons, sentiments and statistical features. The authors of paper [17] use linguistic and syntactic features such as sentence structure, word patterns, word n-grams and part-of-speech (POS) n-grams etc. and learn a classifier using support vector machine (SVM) to detect click-baits. Potthast et. al.[21] use text features and meta-information of tweets such as entity mentions, emotional polarity, tweet length and word n-grams to learn a classifier, experimenting with methods such as random forest, logistic regression etc.to detect click-baits. Yimin et. al.[22] propose to conduct lexical and syntactic analysis and advocate to utilize image features and user-behavior features for the identification of the click-baits. These methods are outperformed by the recent deep learning based methods[8, 13], in which hand crafted feature engineering is not required. News headline incongruence is closely related to a number of tasks such as sentence matching based stance classification [18, 14].

Sentence pair classification task using fine tuned pre-trained language models such as BERT [4] and RoBERTa [5] has received a great traction from the community and it is a closely related problem to headline incongruence. Sentence pair classification typically consists a pair of sentences, while in headline incongruence systems, we need to deal with a sentence and a large news body content in order to form the evidence for congruence. The sentence matching and lexical similarity based methods [20] are not a good fit for headline incongruence problem due to inherent challenges such as relative length and vocabulary mismatch between the news headline text and its body content. Therefore, these tasks share the advancement of the development of the techniques. For example, Wei et. al.[15] introduce a co-training based approach with myriad kinds of features such as sentiments, textual, and informality. Recent works such as [7] use neural attention [1] based approach to achieve headline guided

contextual representation of the news body text and also release a Korean and an English dataset for headline incongruence.

Although some recent works such as [7] have achieved state-of-the-art performance, most of the existing approaches do not perform well in the case of the headline containing cardinal numbers because no additional emphasis is given to the cardinal numbers. [9] use a generative approach to augment the dataset by additionally generating synthetic headlines. Mishra et. al.[3] use an inter-mutual attention-based semantic matching between the original and a synthetically generated headlines via generative adversarial network based techniques, which utilises the difference between all the pairs of word embeddings of words involved and computes mutual attention score matrix. These generative methods are also not very useful in the task at hand as the news headlines, generated using news body content, usually miss the cardinal information. Focusing on the quantity cases, we propose a neural attention mechanism in which, we use novel cardinal POS triplet and cardinal phrase guided attention in addition to standard headline guided attention. This technique makes sure to have two contextual information: firstly, by using headline guided attention, all the keywords of the headline are utilized in forming the overall attention oriented representation. Secondly, by applying cardinal POS triplet and cardinal phrase guided attention, we ensure that cardinal value is emphasized and overall representation contains the effect of cardinal value.

# 3 Problem Definition and Proposed Model

In this section, we formally introduce the problem definition. Then we present the overall architecture of the proposed model *POSHAN* sequentially, see Figure 2. In specific, we first discuss the embedding layer, which outputs the vector representations of the words and cardinal pos-tag patterns. Secondly, we describe the cardinal pos-tag pattern guided hierarchical attention in detail for both word and sentence level. Thirdly, we introduce cardinal phrase guided and headline text guided hierarchical attention. In the end, we explain a method to fuse all the three attention types to get the overall attention scores.

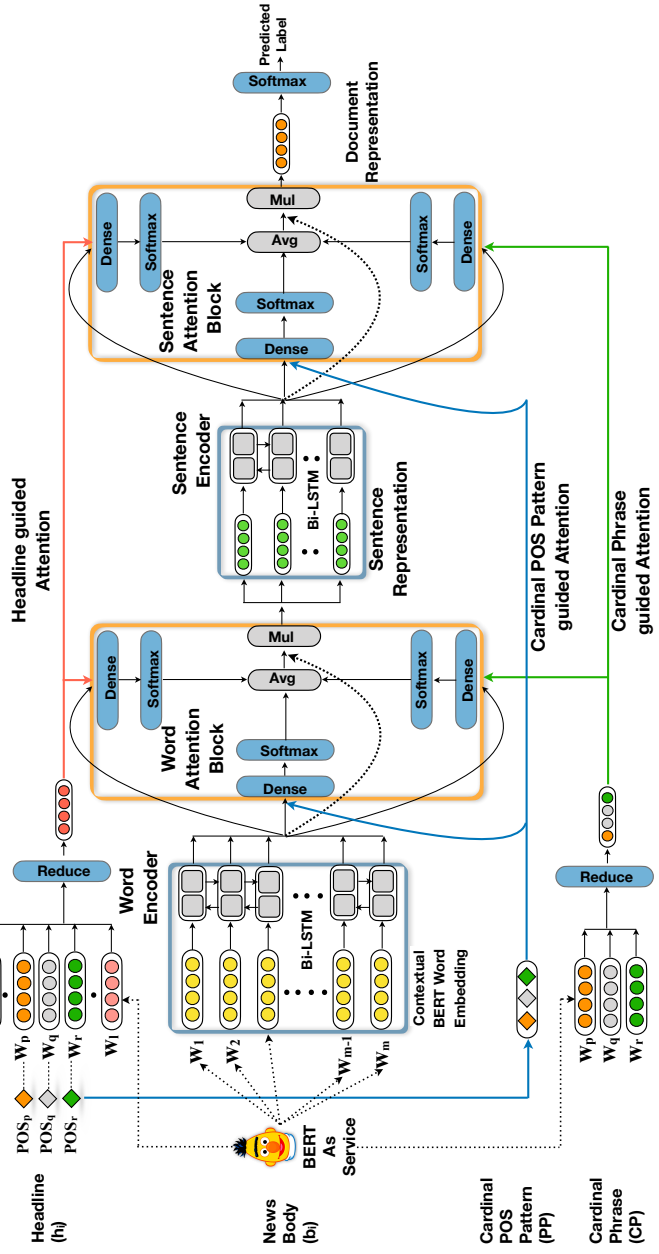



Figure 2: Overall Architecture of POSHAN Model: Rectangle with blue borders are the Bi-LSTM based encoder at the word and sentence levels. Rectangle with orange borders are the Attention blocks at the word and sentence levels. Headline guided Attention,Cardinal POS Pattern guided Attention and Cardinal Phrase guided Attention are depicted as red, blue and green connecting lines respectively.

## 3.1 Problem Definition

Given a news item $n_i \in N$, where $N$ is the set of all news items, which has a headline $h_i$ and body content $b_i$, *news title incongruence detection* aims to predict the news as "Congruent (C)" or "Incongruent (I)', where incongruence denotes a mismatch between $h_i$ and $b_i$ by content. News headline $h_i$ and news body content $b_i$ are comprising of sequence of $l$ words as $h_i = \{w_{h1}, w_{h2}, ..., w_{hl}\} \in W$ and $m$ words as $b_i = \{w_{b1}, w_{b2}, ..., w_{bm}\} \in W$ correspondingly, where $W$ is the overall vocabulary set.

## 3.2 Embedding Layer

In Figure 2, for a news item $n_i$, the corresponding headline $h_i$ of length $l$ and body content $b_i$ of length $m$ are represented as $h_i = \{f(w_{h1}), .., f(w_{hl})\}$ where $f(w_{hj}) \in \mathbb{R}^d$ is a word embedding vector of dimension $d$ for $J^{th}$ word in headline $h_i$ and $b_i = \{f(w_{b1}), f(w_{b2}), ..., f(w_{bm})\}$ where $f(w_{bk}) \in \mathbb{R}^d$ is a word embedding vector of dimension $d$ for $K^{th}$ word in body content $b_i$. We use pre-trained contextual BERT embeddings, extracted using bert-as-service [12] tool to get the embeddings of the size of 768 dimensions for each word. Each headline is associated with a cardinal pos-tag pattern of form $(POS_p\ POS_q\ POS_r)$, where $POS_p$, $POS_q$ and $POS_r$ are the pos-tags corresponding to the cardinal phrase $(W_p\ W_q\ W_r)$. We describe the cardinal pos-tag pattern and cardinal phrase in detail in sections 3.4 and 3.6 respectively. We also create the vector representation $\vec{PP}$ for each of the cardinal pos-tag patterns of the size of 100 dimensions and initialize them with uniformly random weights. We learn weights for these cardinal pos-tag pattern embeddings jointly in the *POSHAN* model via backprop of error, as shown in the Figure 3.

## 3.3 Sequence Encoder

The pre-trained contextual BERT embeddings of the words of the news body text $b_i = \{f(w_{b1}), f(w_{b2}), ..., f(w_{bm})\}$ are fed to a Bi-directional Long short term memory (Bi-LSTM) unit [27] based encoder, which encodes the news body text using standard LSTM equations. The output from Bi-LSTM units are the concatenations of forward and backward

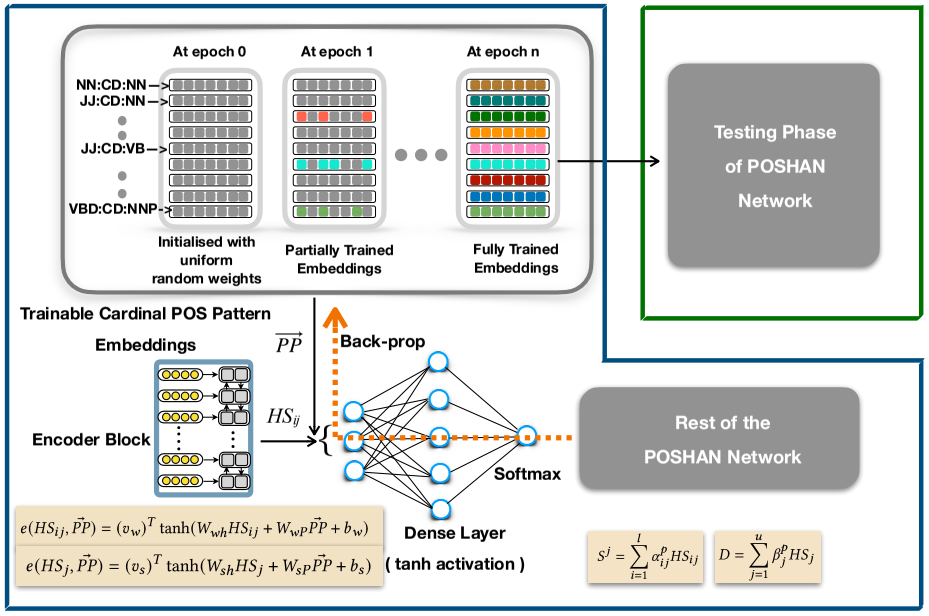



Figure 3: Training of Cardinal POS Pattern Embeddings: Inspired by a recent work [6], we create trainable Cardinal POS Pattern embeddings of 100 dimensions for each POS pattern and initialize them with the uniformly random weights to get the representation of POS patterns in vector space. The weights of these embeddings are trained during training of POSHAN model via back-prop of error. At the test time, we use already trained Cardinal POS Pattern embeddings, trained during training.

hidden states for each word, i.e., $HS_{i,j} = \overrightarrow{hs_{i,j}} \parallel \overleftarrow{hs_{i,j}}$. Where $\overrightarrow{hs_{i,j}}$ and $\overleftarrow{hs_{i,j}}$ are the forward and backward hidden states of Bi-LSTM units. $HS_{i,j}$ is the overall hidden state for the $i^{th}$ word of the $j^{th}$ sentence.

## 3.4 Cardinal POS Triplet Patterns

The idea of utilizing POS-tag Patterns to capture the intended context in natural language text is inspired by prior works [30, 25]. [30] propose and utilize 7 handcrafted part-of-speech (POS) patterns to extract significant and useful phrases from a long unstructured text. We utilize part-of-speech patterns containing cardinal POS tag 'CD' and call it cardinal POS triplet patterns. A cardinal POS triplet pattern can be defined as $(* : CD : *)$, where in place of wildcards, there can be $(JJ, NN, VB)$ etc., e.g. $(NN : CD : JJ)$. In contrast to [30], we do not handcraft a list of the viable POS patterns rather we use the all possible combination of POS patterns of length 3, containing POS tag 'CD'. We apply a neural attention layer in which, these cardinal POS triplets are used to guide the attention to select salient words and sentences which are significant for the POS pattern.

## 3.5 Cardinal POS Triplet Pattern Guided Hierarchical Attention

The objective of the Cardinal POS Triplet Pattern attention is to attend or select salient words that are significant and have some connotation with the cardinal phrase of the headline. Similarly, we aim to attend the salient sentences at the sentence level attention. Yoon have used headline guided attention to model the contextual representation of the news body text. However, we observe that the headline guided attention is not sufficient and effective, in case of headlines containing cardinal values. During experiments, we noticed that only headline-based attention convolutes the effective representation and fails to capture the influence of cardinal phrases on the overall document representation. We take a different and more logical design decision, in which we use part-of-speech patterns contained in each headline $h_i$ to guide the attention. We learn an embedding $\vec{PP}$ for each cardinal POS triplet pattern as discussed in

section 3.2.

**Computing Word Level Attention weights:** We use the embedding of cardinal POS triplet pattern $\vec{PP}$ to compute the attention scores given to each hidden state of the Bi-LSTM encoder.

$$S^j = \sum_{i=1}^{l} \alpha_{ij}^p HS_{ij} \tag{5.1}$$

Where $HS_{ij}$ is the hidden state for the $i^{th}$ word of $j^{th}$ sentence and $l$ is maximum number of words in a sentence. $\alpha_{ij}^p$ is the attention weight. $S^j$ is the formed sentence representation of $j^{th}$ sentence after attention scores are applied. The attention score $\alpha_{ij}$ can be defined as:

$$\alpha_{ij}^p = \frac{\exp(e(HS_{ij}, \vec{PP}))}{\sum_{k=1}^{l} \exp(e(HS_{ik}, \vec{PP}))} \tag{5.2}$$

Where $e$ is a $tanh$ based scoring function, which is used to compute the attention scores. $\vec{PP}$ is the POS-Tag pattern vector. The scoring function $e(HS_{ij}, \vec{PP})$ can be defined as:

$$e(HS_{ij}, \vec{PP}) = (v_w)^T \tanh(W_{wh} HS_{ij} + W_{wP} \vec{PP} + b_w) \tag{5.3}$$

Where $v_w$ is weight vector at the word level. $W_{wh}$ and $W_{wP}$ are the weight matrices for hidden state and aspect vector and $b_w$ is bias at the word level respectively.

**Computing Sentence Level Attention weights:** To compute sentence level POS-Tag pattern driven attention weights, we use POS-Tag pattern vector representation $\vec{PP}$ and hidden states $HS_j^S$ from the sentence level BI-LSTM units as concatenations of both forward and backward hidden states $HS_j^S = \overrightarrow{hs_j^S} \parallel \overleftarrow{hs_j^S}$ as follows:

$$D = \sum_{j=1}^{o} \beta_j^p HS_j \tag{5.4}$$

Where $HS_j$ is the hidden state for $j^{th}$ sentence and $\beta_j^p$ is the attention weight. $o$ is the maximum no of sentences in a news body text. $D$ is the formed document representation after the attention scores are applied. The attention score $\beta_j$ can be defined as:

$$\beta_j^p = \frac{\exp(e(HS_j, \vec{PP}))}{\sum_{k=1}^{o} \exp(e(HS_k, \vec{PP}))} \tag{5.5}$$

Where $e$ is a $tanh$ based scoring function, which is used to compute the attention scores. $\vec{PP}$ is the POS-Tag pattern vector. The scoring function $e(HS_j, \vec{PP})$ can be defined as:

$$e(HS_j, \vec{PP}) = (v_s)^T \tanh(W_{sh} HS_j + W_{sP} \vec{PP} + b_s) \tag{5.6}$$

Where $v_s$ is weight vector at the sentence level. $W_{sh}$ and $W_{sP}$ are the weight matrices for hidden state and aspect vector and $b_s$ is bias at the sentence level respectively.

## 3.6 Cardinal Phrase Guided Hierarchical Attention

We deal with the incongruence detection for the news headlines containing cardinal numbers, therefore the most significant information and cue is the cardinal number itself and neighbouring words. For each headline, we extract a word triplet of form $* : Numerical - value : *$, where in place of wildcards, there can be any words., E.g. $Loan\, 1\, million$. We call these word triplets as cardinal phrases. We use these cardinal phrases to drive attention to select salient words and sentences at word level and sentence level correspondingly. To do that, we represent each cardinal phrase $CP$ as the summation of embeddings of all three words of word triplet as:

$$\vec{CP} = f(W_p) + f(W_q) + f(W_r) \tag{5.7}$$

In a very similar fashion to cardinal POS triplet pattern guided attention, we use $\vec{CP}$ to compute the attention weights at both the word and sentence levels.

$$\alpha_{ij}^{c} = \frac{\exp(e(HS_{ij}, \vec{CP}))}{\sum_{k=1}^{l} \exp(e(HS_{ik}, \vec{CP}))} \tag{5.8}$$

$$\beta_{j}^{c} = \frac{\exp(e(HS_{j}, \vec{CP}))}{\sum_{k=1}^{o} \exp(e(HS_{k}, \vec{CP}))} \tag{5.9}$$

## 3.7 Headline Guided Hierarchical Attention

The objective of the headline driven attention is to select words and sentences in the news body text, which are relevant and topically aligned with headline content. The cardinal POS-tag pattern and cardinal phrase carry useful information regarding cardinal values but to capture the whole context of the headline and it's influence on news body text, we can not get rid of headline driven attention. We represent each headline $\vec{h}$ as the summation of embeddings of all the words contained in it as:

$$\vec{h} = \sum_{x=1}^{l} f(w_x) \tag{5.10}$$

In a very similar fashion to cardinal POS triplet pattern guided attention, we use $\vec{h}$ to compute the attention weights at both the word and sentence levels.

$$\alpha_{ij}^{h} = \frac{\exp(e(HS_{ij}, \vec{h}))}{\sum_{k=1}^{l} \exp(e(HS_{ik}, \vec{h}))} \tag{5.11}$$

$$\beta_{j}^{h} = \frac{\exp(e(HS_{j}, \vec{h}))}{\sum_{k=1}^{o} \exp(e(HS_{k}, \vec{h}))} \tag{5.12}$$

## 3.8 Fusion of Attention Weights and Classification

We compute the overall attention weights from three kinds of attention mechanisms: POS-pattern-driven, Cardinal-phrase-driven, and headline driven attention at both the word and sentence levels. At the word level:

Table 5.1: Gist of attention schemes

| Attention | Hierarchical | Driven by | Text based | Linguistic based |
|---|---|---|---|---|
| Cardinal POS Pattern | ✓ | POS tag triplet | | ✓ |
| Cardinal Phrase | ✓ | Word triplet | ✓ | |
| Headline | ✓ | All words of headline | ✓ | |

$$\alpha_{i,j} = (\alpha^p_{i,j} + \alpha^c_{i,j} + \alpha^h_{i,j})/3 \tag{5.13}$$

$$S^j = \sum_{i=1}^{l} \alpha_{ij} HS_{ij} \tag{5.14}$$

where $\alpha^p_{i,j}$, $\alpha^c_{i,j}$ and $\alpha^h_{i,j}$ are the attention weight vectors from POS-pattern, Cardinal-phrase and headline-attention at the word level. $S^j$ is the formed sentence representation after overall attention for the $j^{th}$ sentence. At the sentence level:

$$\beta_j = (\beta^p_j + \beta^c_j + \beta^h_j)/3 \tag{5.15}$$

$$D = \sum_{j=1}^{o} \beta j HS_j \tag{5.16}$$

where $\beta^p_j$, $\beta^c_j$, and $\beta^h_j$ are the attention weight vectors from POS-pattern, Cardinal-phrase and headline-attention at the sentence level, and $D$ is the formed document representation after overall attention. The document representation $D$ is used with a Softmax layer with softmax cross-entropy with logits as loss function for the classification. We compute the predicted label $\hat{y}$ as:

$$\hat{y} = softmax(W_{cl} D + b_{cl}) \tag{5.17}$$

Where $W_{cl}$ and $b_{cl}$ are the weight matrix and bias term.

# 4 Dataset Creation

For evaluation, we create the datasets driven by two publicly available datasets, NELA17 and Click-bait Challenge[17] (cf. Table 5.2). Yoon

[17] `http://www.clickbait-challenge.org/`

provide a script[18] to create the NELA17 dataset from an original news collection NELA17 dataset[19]. The NELA17 dataset comprises of 45521 congruent and 4551 incongruent news headline-body pairs. The Click-bait Challenge dataset is created via crowd-sourcing based annotation of a collection of social media posts. The Click-bait Challenge dataset contains 16150 and 4883 social media posts, which are annotated as congruent and incongruent correspondingly. Using NELA17 and Click-bait Challenge datasets, we derive two new datasets, in which all the news headlines in NELA17 and all the social media posts in Click-bait Challenge, contain a numerical value. We call these two new datasets as Derived NELA17 dataset and Derived Click-bait Challenge dataset. We create the new datasets using these steps:

(1) We use POS tagger to get the words in headlines tagged with one of the corresponding Penn Treebank POS Tag Set.

(2) We keep all the headline-body pairs in which pos-tag CD (cardinal) appears in headline.

The statistics of these datasets are reported in Table 5.3. We also extract two new features for each headline-body pair:

- A pos-tag triplet of form $* : CD : *$, where in place of stars, there can be JJ, NN, VB etc. E.g. $NN : CD : JJ$. We call this pos-tag triplet as Cardinal POS-tag Pattern. For a vector representation for each of the cardinal pos-tag pattern, we create a trainable embeddings of the size of 100 dimensions and initialize them with uniformly random weights. The weights for these embeddings are learned jointly using hierarchical attention in the *POSHAN* model.

- A word triplet of form $* : Numerical - value : *$, where in place of stars, there can be any words. E.g. $Loan\ 1\ million$. We call this word triplet as Cardinal Phrase.

[18] https://github.com/sugoiii/detecting-incongruity-dataset-gen

[19] https://github.com/BenjaminDHorne/NELA2017-Dataset-v1

Table 5.2: Dataset Statistics for NELA17

| **Statistics** | **NELA17** | **Derived NELA17** |
|---|---|---|
| Incongruent | 45521 | 6234 |
| Congruent | 45521 | 7766 |
| Total | 91042 | 14000 |

# 5 Experimental Evaluation

## 5.1 Experimental Details

### 5.1.1 Baselines

We compare our model with the following baselines:

**SVM [29]** : We start with feature-based methods utilizing support vector machine (SVM) by considering both linguistic and statistical features. In specific, we use word tri-grams, four-grams, and part-of-speech bi-grams and tri-grams as features to learn a classifier using SVM. Usually, a click-bait headline contains word phrases like "what happens if" and "You will Never Believe", which can be easily captured by tri-grams and four-grams. Besides, POS tag combinations such as "PRP WD RB" are more frequent in incongruent headlines than incongruent ones, therefore, part-of-speech bi-grams and tri-grams are used to learn this distinguishing feature.

**LSTM [27]** : We use long short term memory unit to encode both headline and body pair and apply softmax for the classification. We use pretrained GloVe embeddings of size 100 and the size of the hidden states of the Bi-LSTM unit is kept at 200. The concatenation of the news headline and the body text is used as input to the Bi-LSTM based encoder.

**POSAt** : We propose a baseline method called as POS-tag guided Attention (POSAt) and compare the performance of our proposed approach *POSHAN*, as this method uses POS tags to give importance to certain words. This method is inspired by a recent work [11]. We use NLTK POS tagger to tag each word in the news headline/body pairs and maintain

Table 5.3: Dataset Statistics for Click-bait Challenge

| Statistics | Click-bait Challenge | Derived Click-bait Challenge |
|---|---|---|
| Incongruent | 4883 | 754 |
| Congruent | 16150 | 2681 |
| Total | 21033 | 3435 |

the mapping between words and corresponding POS tags using an index. POS tags are categorized into 6 semantic categories for sake of simplicity and brevity.

(1) **Noun chunk:** NN, NNS, NNP, NNPS

(2) **Verb chunk:** VB, VBD, VBG, VBN, VBP, VBZ

(3) **Adjective chunk:** JJ, JJR, JJS

(4) **Pronoun chunk:** WP, WP

(5) **Adverb chunk:** WRB

(6) **Cardinal numbers chunk:** CD

The POS tags for each word is represented as a 6-dimensional vecotr $g(x_i) \in \mathbb{R}^6$ of length 6. The weights of these embedding vectors are initialized in two ways.

(1) **Initialize with very less value, close to zero:** In case of all zeros initialization, model performs well.

(2) **Initialize with random weights:** In case of random weight initialization performance of the model degrades.

These POS tag vector sequences are fed into a fully connected POS tag embeddings layer so that their weights are also trainable. Each part-of-speech category is assigned with one attention weight $\theta_i$, which will be learned during training. In this way, each word is represented by $f'(w_i) = f(w_i) \times \theta_i$. We train a small neural network with only single hidden layer to learn weights for each POS tag category and then we use a custom lambda layer to reshape the POS weight tensor into a compatible shape so

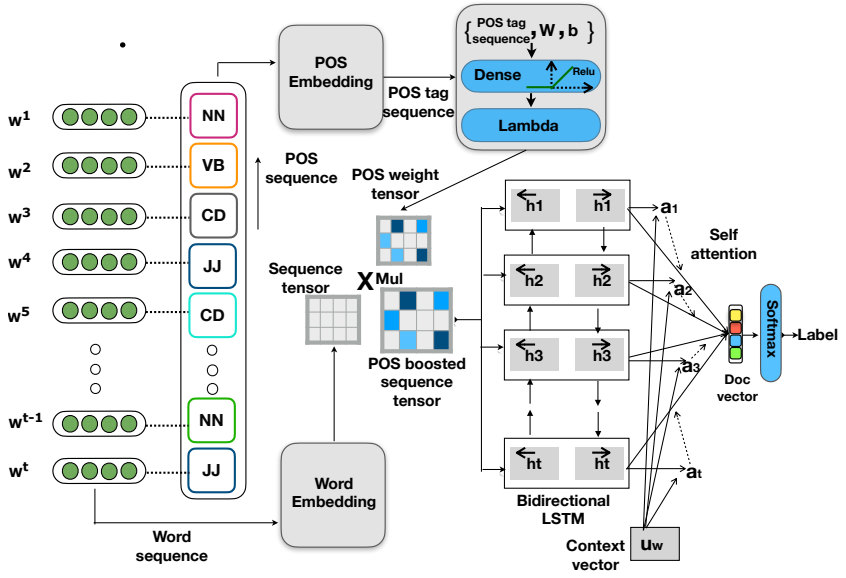



Figure 4: Depiction of the POS tag guided attention model. From left to right, the sequence of words of the news item and corresponding part-of-speech tags go through word embedding and POS embedding layers. Attention scores are computed for each POS tag via a dense layer with ReLu activation. The resultant weight score matrix is multiplied to the word embedding sequence matrix. Attended representation is fed to a Bi-LSTM unit. Lastly, the resultant document representation from Bi-LSTM is used with a softmax layer for classification.

that we can boost each word vector with it's corresponding POS weight vector.

In very similar fashion to LSTM baseline, pretrained GloVe embeddings of size 100 are used to represent the word vectors and hidden states of the LSTM unit is kept at 200. The nltk library[20]with MaxEnt POS Tagger [28] is used to tag the concatenation of the news headline and the body text.

[20]https://pypi.org/project/nltk/

**Yoon [7]** : This is a state of the art method for news headline incongruence detection. It uses a hierarchical dual encoder based model which uses headline guided attention to learn the contextual representation. The original [7] paper uses a Korean news collection as dataset for evaluation but they also release an English version of the dataset called as NELA17. We use their model with NELA17 dataset, keeping all the settings as prescribed in [7].

**BERT-Sent_Pair [4]** : We fine tune a pretrained BERT model for sequence pair classification task. We utilize the Hugging face transformers and dataset libraries to download pre-trained model. We use pre-built "BertForSequenceClassification", provided by Hugging face library. Headlines and body pairs are packed together into a single sequence with adequate padding.

**MuSem [3]** : This is a very recent work related to title incongruence detection, which uses both the NELA17 and Click-bait challenge datasets for evaluation. The authors propose a method that uses inter-mutual attention-based semantic matching between the original and a synthetically generated headlines via generative adversarial network based techniques, which utilises the difference between all the pairs of word embeddings of words involved and computes mutual attention score matrix.

## 5.2 *POSHAN* Implementation Details

The *POSHAN* model is implemented [21] using TENSORFLOW 1.10.0[22] platform. For performance evaluation, Macro F1, and Area Under the ROC Curve (AUC) scores are used as performance metrics. We keep the size of hidden states of bi-directional Long Short-term Units (LSTM) as 300, the size of embedding dimensions of pretrained BERT [4] embeddings as 768. We use softmax cross-entropy with logits as the loss function. We keep the learning rate as 0.003, batch size as 128, and gradient clipping as 6. The parameters are tuned using a grid search. We use $50$ epochs for each model and apply early stopping if validation loss

[21] https://github.com/rahulOmishra/POSHAN_CIKM
[22] https://www.tensorflow.org/install/source

Table 5.4: Comparison of the proposed model *POSHAN* with various state-of-the-art baseline models for NELA17 Dataset. The results for *POSHAN* are statistically significant ($p - value = 1.32e^{-2}$ for NELA17 Dataset using pairwise student's t-test

| Derived NELA17 Dataset | | |
|---|---|---|
| **Model** | **Macro F1** | **AUC.** |
| SVM [29] | 0.608 | 0.610 |
| LSTM[27] | 0.627 | 0.639 |
| POSAt | 0.624 | 0.637 |
| BERT-Sent_Pair [4] | 0.642 | 0.658 |
| Yoon [7] | 0.653 | 0.659 |
| MuSeM[3] | 0.703 | 0.721 |
| POSHAN | **0.748** | **0.763** |
| Original NELA17 Dataset | | |
| **Model** | **Macro F1** | **AUC.** |
| SVM [29] | 0.622 | 0.637 |
| LSTM [27] | 0.642 | 0.663 |
| POSAt | 0.648 | 0.669 |
| BERT-Sent_Pair[4] | 0.677 | 0.683 |
| Yoon [7] | 0.685 | 0.697 |
| MuSeM[3] | 0.752 | 0.769 |
| POSHAN | **0.765** | **0.783** |

does not change for more than 5 epochs. We keep maximum words in a sentence as 45 and maximum number of sentences in a news body text as 35.

**Handling the multiple cardinal values** : There are some cases where news headlines contain multiple cardinal values such as in fig 1. At the training time, to utilize the context of all cardinal values present in the headline, we replicate the news headline and body pair for each cardinal value in train set. At the test time however, we concatenate all the learned

Table 5.5: Comparison of the proposed model *POSHAN* with various state-of-the-art baseline models for click-bait challenge dataset. The results for *POSHAN* are statistically significant ($p - value = 2.29e^{-3}$ for click-bait challenge dataset using pairwise student's t-test.

| Derived Click-bait challenge Dataset | | |
|---|---|---|
| **Model** | **Macro F1** | **AUC.** |
| SVM [29] | 0.596 | 0.608 |
| LSTM [27] | 0.604 | 0.617 |
| POSAt | 0.614 | 0.620 |
| BERT-Sent_Pair[4] | 0.637 | 0.649 |
| Yoon [7] | 0.646 | 0.659 |
| MuSeM[3] | 0.698 | 0.717 |
| POSHAN | **0.739** | **0.748** |
| Click-bait challenge Dataset | | |
| **Model** | **Macro F1** | **AUC.** |
| SVM [29] | 0.618 | 0.629 |
| LSTM [27] | 0.630 | 0.641 |
| POSAt | 0.636 | 0.649 |
| BERT-Sent_Pair[4] | 0.653 | 0.662 |
| Yoon [7] | 0.660 | 0.678 |
| MuSeM[3] | 0.735 | 0.747 |
| POSHAN | **0.743** | **0.761** |

cardinal POS tag vectors pertaining to the same news headline and use this overall POS tag vector to guide the attention.

**Extraction of BERT Embeddings [12]** : We use bert-as-service, which utilizes extract_features.py file from original BERT implementation, to extract the word embeddings from pretrained BERT model. We fine tune uncased_L-12_H-768_A-12 pretrained BERT model for sentence pair classification task. We set pooling_strategy argument to NONE and use our own tokenizer. We use fine tuned BERT model to extract embeddings of 768 dimensions for each word.

## 5.3 Results

In this section, we compare the results of the *POSHAN* model with the baselines and state-of-the-art methods.

### 5.3.1 Results for NELA17 Dataset

In Table 5.4, we observe that in case of Derived NELA17 dataset, all the deep learning based methods outperform the non-deep learning method such as SVM model, which uses linguistic features and gets $0.608$ and $0.610$ in terms of Macro F1 and AUC. The POSAt model with Macro F1 score as $0.624$ and AUC as $0.637$, performs comparable with vanilla LSTM model. In our experiments, we introspect that the design decision in POSAt model to apply POS-tag guided attention at the POS-tag chunk level, does not result in effective representation and provides very less intended effects of POS types on words. This way of POS-tag guided attention learns the attention score at the POS category level only as discussed in Section 5.1.1 such as Noun chunk, Verb chunk etc.

The BERT-Sent_Pair with Macro F1 as $0.642$ and AUC as $0.658$, outperforms the POSAt model with significant difference. This gain can be attributed to the better contextual representation of words, learned in form of transformer based BERT embeddings. On the other hand, Yoon model [7] performs slightly better than BERT-Sent_Pair with Macro F1 as $0.653$ and AUC as $0.659$. In addition to hierarchical encoder, which captures the complex structure of the news body content, having inherent hierarchical

nature, Yoon model also uses a headline driven hierarchical attention, which not only selects salient and relevant words and sentences but also reduces the effective length of the news body. In contrast, vanilla LSTM, POSAt and even BERT-Sent_Pair model did not scale well for long text sequences. The MuSem model uses generative adversarial network based synthetic headline generation methods to generate a very low dimensional headline corresponding to news body and applies a novel mutual attention based semantic matching for incongruence detection. The MuSem model achieves significant gains over Yoon model, due to low dimensional representation of news body and effective semantic matching technique.

The proposed *POSHAN* model beats all the other methods achieving $0.748$ and $0.763$ as Macro F1 and AUC, respectively. The potential reason behind this better performance is superior document representation learned due to proposed attention mechanisms, which give adequate importance to significant cardinal values present in headline. In contrast, both Yoon and MuSem models fail to capture cues pertaining to cardinal patterns and phrases.

In case of Original NELA17 Dataset, we notice a very similar trend as with original NELA17 dataset, however performance of all the models improved with a significant margin. On the other hand, *POSHAN* happens to yield more improvement in performance compared to other models for original dataset as a bonus. These gains can be attributed to cardinal POS-tag pattern based attention and cardinal phrase guided attention in addition to headline guided attention significantly.

### 5.3.2 Results for Click-bait Challenge Dataset

In case of both the derived and original click-bait challenge dataset also, we see a very similar performance chart. All the deep learning based methods outperform the non-deep learning method such as SVM model With $0.596$ and $0.608$ in Macro F1 and AUC. The MuSem model with $0.698$ and $0.717$ in Macro F1 and AUC, outperforms all the other baselines. The proposed model *POSHAN* performs better than MuSem model with significant gains and these gains can be explained by very similar reasoning, as provided in Section 5.3.1.

## 5.4 Ablation Study

In Table 5.6, we report an ablation study of *POSHAN* using Derived NELA 17 Dataset. We used derived dataset for ablation study rather than original dataset because we want to asses the importance of different components of the *POSHAN* with major focus of this paper, which is news headlines with important numerical values. In the ablation version 1), we remove the cardinal POS-tag pattern guided attention and keep the other two methods of attention intact and this step results in significant decrease in performance, which proves the usefulness and effectiveness of the cardinal POS-tag pattern guided attention. This corroborates with our original hypothesis and intuition. In the ablation version 2), we remove cardinal phrase attention and observe very similar decrease in performance. In the ablation version 3), we replace headline guided attention with headline encoder, in which we encode the words of news headlines in addition to news body words and concatenate the overall encoded sequence. We observe that without headline guided attention, model performs poorly because just a simple concatenation of encoded body and headline word sequences does not result in contextually important representation.

We can conclude from 1), 2) and 3), that although all the three attention mechanism are effective individually too but combination of all the three becomes more effective. In the ablation version 4), we replace the pre-trained BERT embbedings with GloVe [24], due to which the performance degrades drastically. The reason behind such a drop in the results is that the BERT embeddings provide superior contextual information than GloVe pretrained embeddings. We do not see much change in results in ablation version 5) as Bi-GRU [23] and Bi-LSTM perform pretty much the same with our dataset. The performance of the model decreases a bit with replacement of Bi-LSTM with LSTM units in ablation version 6) and the obvious reason behind this better context learned by Bi-LSTM compared to LSTM units.

## 5.5 Error Analysis

We conduct an error analysis of MuSem [3] and *POSHAN* model with Derived Click-bait challenge dataset in Table 5.7. In the case of MuSem model, we observe 235 false negatives(FN) and 201 false positives (FP),

Table 5.6: Ablation study of POSHAN and Yoon[7] model conducted on derived NELA 17 Dataset.

| Derived NELA 17 Dataset | | |
|---|---|---|
| **Scenario** | **Macro F1** | **AUC.** |
| Original POSHAN | 0.748 | 0.763 |
| 1) Remove Cardinal POS Att | 0.726 | 0.742 |
| 2) Remove Cardinal Phrase Att | 0.731 | 0.749 |
| 3) Replace Headline Att with Headline Enc | 0.648 | 0.669 |
| 4) Replace BERT with Glove | 0.716 | 0.736 |
| 5) Replace Bi-LSTM with Bi-GRU | 0.746 | 0.761 |
| 6) Replace Bi-LSTM with LSTM | 0.741 | 0.759 |
| Derived NELA 17 Dataset | | |
| **Scenario** | **Macro F1** | **AUC.** |
| Original Yoon | 0.653 | 0.659 |
| 1) Remove Headline Att | 0.593 | 0.595 |
| 2) Replace para to sent level Att | 0.649 | 0.653 |
| 3) Replace Glove with W2V | 0.610 | 0.618 |
| 4) Replace Bi-LSTM with Bi-GRU | 0.652 | 0.657 |
| 5) Replace Bi-LSTM with LSTM | 0.641 | 0.648 |

on the other hand, *POSHAN* produces 207 false negatives and 179 false positives. We notice that the major improvement with *POSHAN* model, occurs in false negatives from 235 to 207, and most of these incorrectly predicted samples were related to important cardinal figures mentioned in the news headlines such as 'Indiana couple admits to stealing 1.2 Million dollars from Amazon'. We also observe some incorrectly predicted false-positive cases by *POSHAN* model because of the wrong POS-tag

Table 5.7: Error Analysis of Yoon model [7] and *POSHAN* with Derived Click-bait challenge Dataset

| **Model** | **False Positives** | **False Negatives** |
|---|---|---|
| MuSem | 201 | 235 |
| *POSHAN* | 179 | 207 |

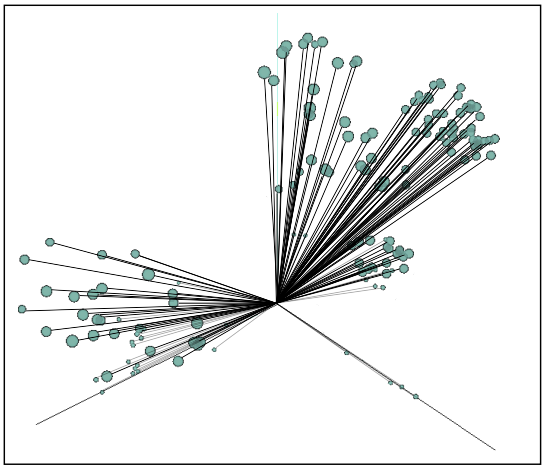


Figure 5: Visualization of Cardinal POS Pattern Embeddings

assignment by POS tagger and due to this *POSHAN* misses out on the opportunity to consider those cardinal POS patterns and cardinal phrases.

## 5.6 Visualization of Cardinal POS Pattern Embeddings

In the Figure 5, we present a visualization of cardinal pos-tag patterns. To visualize the learned embeddings of the cardinal pos-tag patterns, we use t-Distributed Stochastic Neighbor Embedding (t-SNE) [26] with parameters as perplexity = 10, learning rate = 0.1 and iterations = 1000. The t-SNE method produces the visualization in a low dimensional space. We observe that the Cardinal POS Patterns have formed clearly separated clusters

in embedding space, which connotes the congruence and incongruence labels. We also observe that cardinal POS patterns with similar tags such as $(NN : CD : JJ)$ and $(NNS : CD : JJ)$ are closer in embedding space and on the other hand the patterns with disjoint tag combinations such as $(NN : CD : JJ)$ and $(VBG : CD : CD)$ are farther apart from each other.

## 5.7 Visualization of Attention Weights

**Headline:-** Immigration Expert : US Will Have 100 Million New Immigrants in Next 50 Years.

**Yoon Model**
These projections show that new immigrants and their descendants will drive. most U.S. population growth in the coming 50 years, as they have for the, past half-century. Among the. projected 441 million Americans in 2065, 78 million will be: immigrants and 81 million will be people born in U.S. to immigrant parents.

**POSHAN Model**
These projections show that new immigrants and their descendants will drive. most U.S. population growth in the coming 50 years, as they have for the, past half-century. Among the projected 441 million Americans in 2065, 78 million will be: immigrants and 81 million? will be people born in U.S. to immigrant parents.

Figure 6: Attention weight visualization: Word level attention weights from Yoon Model and POSHAN model for an anecdotal example are presented by highlighting the individual words (Best viewed in color). The depth of the color represents the strength of the attention weights.

In Figure 6, to analyse the interpretability of our model *POSHAN* and to showcase the effectiveness of the proposed attention mechanism in forming the contextually important representations, we visualize the attention maps and compare it with Yoon model. In Figure 6, we use distribution of word level attention weights learned from both *POSHAN* and Yoon

model for an anecdotal example by highlighting the individual words. The depth of the color highlights represents the distribution of attention weights. Despite of common headline driven attention in both the models, we observe some clear differences between attention maps of Yoon model and *POSHAN* model due to additional cardinal pos-tag pattern and cardinal phrase guided attention mechanisms in *POSHAN* model. The Yoon model successfully attends some words such as 'immigrants', 'growth' and 'projections' etc. relevant to headline context but fails to capture any words pertaining to significant cardinal phrases such as '78 million' and '50 years' etc. On the other hand, *POSHAN* model not only gives the importance to the words captured by Yoon model but also, it focuses on important cardinal phrases, which is in concert with our intuition about modeling the POS-tag pattern and cardinal phrase based attention.

# 6 Conclusions and Future Work

In this paper, we introduce a novel task of incongruence detection in the news when the news headline contains significant cardinal values. The existing methods fare poorly as they fail to capture the context, pertaining to cardinal values. We present a joint neural model *POSHAN*, which uses three kinds of hierarchical attention mechanisms, namely cardinal POS-tag pattern guided, cardinal phrase guided and news headline guided attention. In the ablation study, we found that cardinal POS-tag pattern guided attention is very significant and effective in forming the cardinal quantity informed document representation. In the evaluation with two publicly available datasets, we notice that *POSHAN* outperforms all the baselines and state-of-the-art methods. Visualization of cardinal POS-tag pattern embeddings and overall attention weights establish the effectiveness of the proposed model, decipher the model's decisions and make it more interpretable and transparent.

In the future, we plan to model the degree of importance of cardinal values in news headlines and also we envisage an assessment of the applicability of the proposed model in case of textual entailment and fact verification tasks such as FEVER [10] dataset, in presence of cardinal values.

# References

[1] **Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio**. *Neural Machine Translation by Jointly Learning to Align and Translate*. In ICLR 2015 as oral presentation.

[2] **Julio Reis, Pedro Olmo, Raquel Prates, Haewoon Kwak, and Jisun An**. "Breaking the News : First Impressions Matter on Online News." In: (), pp. 357–366.

[3] **Rahul Mishra, Piyush Yadav, Remi Calizzano, and Markus Leippold**. "MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching." In: *International Conference on Machine Learning and Applications (ICMLA) 2020*. Miami, Florida, Dec. 2020.

[4] **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[5] **Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov**. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[6] **Rahul Mishra and Vinay Setty**. "SADHAN: Hierarchical Attention Networks to Learn Latent Aspect Embeddings for Fake News Detection." In: ICTIR '19. Santa Clara, CA, USA, 2019, pp. 197–204. ISBN: 9781450368810.

[7] **Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung**. "Detecting Incongruity between News Headline and Body Text via a Deep Hierarchical Encoder." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 791–800.

[8] **Vaibhav Kumar, Dhruv Khattar, Siddhartha Gairola, Yash Kumar Lal, and Vasudeva Varma**. "Identifying Clickbait: A Multi-Strategy Approach Using Neural Networks." In: *The 41st International ACM SIGIR Conference on Research and Development*

*in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, pp. 1225–1228. ISBN: 9781450356572. DOI: 10.1145/3209978.3210144. URL: https://doi.org/10.1145/3209978.3210144.

[9] **K. Shu, S. Wang, T. Le, D. Lee, and H. Liu**. "Deep Headline Generation for Clickbait Detection." In: *2018 IEEE International Conference on Data Mining (ICDM)*. 2018, pp. 467–476.

[10] **James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal**. "FEVER: a Large-scale Dataset for Fact Extraction and VERification." In: New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074.

[11] **Z. Wang, X. Liu, L. Wang, Y. Qiao, X. Xie, and C. Fowlkes**. "Structured Triplet Learning with POS-Tag Guided Attention for Visual Question Answering." In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 1888–1896.

[12] **Han Xiao**. *bert-as-service*. https://github.com/hanxiao/bert-as-service. 2018.

[13] **Ankesh Anand, Tanmoy Chakraborty, and Noseong Park**. "We Used Neural Networks to Detect Clickbaits: You Won't Believe What Happened Next!" In: *Advances in Information Retrieval*. Springer International Publishing, 2017, pp. 541–547. ISBN: 978-3-319-56608-5.

[14] **Zhiguo Wang, Wael Hamza, and Radu Florian**. "Bilateral Multi-Perspective Matching for Natural Language Sentences." In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 4144–4150. DOI: 10.24963/ijcai.2017/579. URL: https://doi.org/10.24963/ijcai.2017/579.

[15] **Wei Wei and Xiaojun Wan**. "Learning to Identify Ambiguous and Misleading News Headlines." In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 4172–4178. ISBN: 9780999241103.

[16] **Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer**. ""8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality." In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, 2016, pp. 94–100.

[17] **A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly**. "Stop Clickbait: Detecting and preventing clickbaits in online news media." In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016, pp. 9–16.

[18] **William Ferreira and Andreas Vlachos**. "Emergent: a novel dataset for stance classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1163–1168. DOI: 10.18653/v1/N16-1138. URL: https://www.aclweb.org/anthology/N16-1138.

[19] **Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout**. "Social Clicks: What and Who Gets Read on Twitter?" In: *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. SIGMETRICS '16. Antibes Juan-les-Pins, France: Association for Computing Machinery, 2016, pp. 179–192. ISBN: 9781450342667. DOI: 10.1145/2896377.2901462.

[20] **Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru**. "Learning Text Similarity with Siamese Recurrent Networks." In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 148–157. DOI: 10.18653/v1/W16-1617.

[21] **Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen**. "Clickbait Detection." In: *ECIR'16*. ECIR'16 1 (2016).

[22] **Yimin Chen, Niall J. Conroy, and Victoria L. Rubin**. "Misleading Online Content: Recognizing Clickbait as "False News"." In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. WMDD '15. Seattle, Washington, USA: Associa-

tion for Computing Machinery, 2015, pp. 15–19. DOI: 10.1145/2823465.2823467.

[23] **Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio**. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL].

[24] **Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean**. "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 3111–3119. URL: http://papers.nips.cc/.

[25] **Rahul Potharaju, Navendu Jain, and Cristina Nita-Rotaru**. "Juggling the Jigsaw: Towards Automated Problem Inference from Network Trouble Tickets." In: *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. Lombard, IL: USENIX Association, Apr. 2013, pp. 127–141. ISBN: 978-1-931971-00-3.

[26] **Laurens van der Maaten and Geoffrey Hinton**. *Visualizing data using t-SNE*. 2008.

[27] **Sepp Hochreiter and Jürgen Schmidhuber**. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667.

[28] **Adwait Ratnaparkhi**. "A Maximum Entropy Model for Part-Of-Speech Tagging." In: *Conference on Empirical Methods in Natural Language Processing*. 1996. URL: https://www.aclweb.org/anthology/W96-0213.

[29] **Corinna Cortes and Vladimir Vapnik**. "Support-Vector Networks." In: *Mach. Learn.* 20.3 (Sept. 1995), pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411.

[30] **John S. Justeson and Slava M. Katz**. "Technical terminology: some linguistic properties and an algorithm for identification in text." In: *Natural Language Engineering* 1.1 (1995), pp. 9–27. DOI: 10.1017/S1351324900000048.