Universitetet i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: Computer Science	Spring / Autumn semester, 20..... Open/Confidential
Author: Priyanka Chennam Lakshmikumar	..... (signature of author)
Programme coordinator: Supervisor(s): Vinay Jayarama Setty	
Title of master's thesis: Fake News Detection - A Deep Neural Network	
Credits: 30	
Keywords: Deep Learning, Machine Learning, Fake News Detection	Number of pages: ..54..... + supplemental material/other: ..... Stavanger,..15 June 2019..... date/year

University of Stavanger

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

Fake News Detection-A Deep Neural Network

Master's Thesis in Computer Science

by

Priyanka Chennam Lakshmikumar

Internal Supervisors

Vinay Jayarama Setty

Rahul Mishra

Reviewers

Vinay Jayarama Setty

Rahul Mishra

June 15, 2019

“Be the change that you wish to see in the world ”

Mahatma Gandhi

Abstract

News is an important source of information for people.Identifying the inaccurate news is a difficult problem.

Fake news, defined by the New York Times "as a made-up story with an intention to deceive " , often for a secondary gain, is arguably one of the most serious challenges facing the news industry today [1].

In this world of busy schedules, none of us have enough time to verify the source of all the news articles. This is a difficult and time taking procedure. Of all the challenges in the widespread of fake news, different types of fake news are a major challenge.

Fake news is a major threat which is basically formed by all forms of false, inaccurate or misleading information

The goal of the Fake News Detection is to explore how machine learning and natural language processing, might be useful to address the fake news problem.

One application that we will explore in this paper is to use Machine Learning techniques to determine whether a pair of news statements agrees, disagrees, or is unrelated to another i.e

The input would be a corpus of news statement pairs and the output will be the classification of the statements on the basis of whether they both agree, disagree, or are unrelated .

We developed several deep neural network-based models to tackle the fake news detection problem, ranging from relatively simple feed-forward networks to elaborate models featuring attention and multiple feature fusion technique using pre-trained GloVe word embeddings.

Our approach learns two representations, one based on attention [2] to decompose the problem into subproblems that can be solved separately, thus making it trivially parallelizable and another based on sequential composition (LSTM) [3] with pre-trained GloVe word embeddings [4]. Both views are used for prediction.

Our models are shown to be effective in efficiently classifying the corpus of statement pairs in a set of experiments using the ByteDance dataset

Acknowledgements

My accomplishment of this thesis is a result of support and guidance from many people around me.

I would like to express my earnest gratitude to Mr.Vinay Jayarama Setty, my supervisor, who was instrumental in introducing me to this topic and guiding me throughout the thesis.

I would like to specially thank Mr.Rahul Mishra, my mentor at University of Stavanger , for supporting and engaging me in a interesting topic for the thesis and conversing all the ideas and helping me spearhead in the right direction of my thesis.

I am grateful to both of them for providing me this wonderful opportunity to pursue this thesis under their supervision and amidst their busy schedule for providing me insightful and valuable feedback .

Finally, I would like to acknowledge with gratitude, the support of my parents ,my brother, my husband Mr.Suraj Naidu and my kids Ahanaa and Anika, for always being my pillars of strength and showing huge trust and confidence in me which kept me going forward.

Priyanka Chennam Lakshmikumar

University Of Stavanger.

Contents

Abbreviations

CNN	**C**onvolutional **N**eural **N**etworks
RNN	**R**ecurrent **N**eural **N**etworks
DNN	**D**eep **N**eural **N**etworks
LSTM	**L**ong **S**hort **T**erm **M**emory
CBOW	**C**ontinuous **B**ag **O**f **W**ords
TF	**T**erm **F**requency
NN	**N**eural **N**etwork
IDF	**I**nverse **D**ocument **F**requency

Chapter 1

Introduction

In olden days, common man used to wait for the next day to see what has happened in the world yesterday. In todays world this is entirely different, news travels almost at the speed of light.

In this world of advanced technology, there is also high competition among-st people as who can generate more news which has led to the birth of fabricated news also known as Fake News.

In the recent past, a newly framed word called "Fake News" has evolved which represents typically fabricated news or false hype comprising information broadcasted through various forms of media.

The main purpose to disperse such news is to misguide or misinform the common reader, damage fame of any person, firm or a nation, create confusion, gain financially or to gain politically from sensationalism.

For e.g. in Germany during the political campaign in 2016, Chancellor of Germany Ms Angela Merkel specifically called out on the issue of false news reports immediately after her election as a leader for fourth term. She emphasized to the public/ government bodies on Internet trolls, bots, and fake news websites. [5].

These fake news information were becoming a force in increasing the power of populist extremism and has announced measures to be regulated in the future. One such instance which was clearly bought to the public notice was on the award-winning German journalist Claas Relotius who resigned from Der Spiegel in 2018 after he himself admitting various cases of journalistic fraud.

This fake news is always being presented as factually accurate however in reality it is not. In todays world we believe what we see on the websites or social media and do not try and pursue to validate if the provided information is true or false.

Since people are often unable to spend enough time to cross-check reference and be sure of the credibility of news, intelligent detection of fake news is essential. Therefore, it is receiving high attention from the research community all over the world.

In countries like Singapore, there is an existing stern action on people who are involved with fake news. Singapore criminalizes the propagation under existing law and says that any individual who broadcasts a message which he knows to be not true or constructed news shall be guilty of a crime.

In Singapore, Google and Facebook have opposed the introduction of new laws to combat fake news, claiming that existing legislation is sufficient to address the problem and that an effective way of fighting fake news is through coaching people on how to differentiate from fake news vs real news.

Despite all these efforts done by the existing society, people, technology and processes we still see fake news in some shape or form every day. [5].

One more classic example where fake news becoming more dominant and creating a wave among people and making sure that the original news is dead completely is from India. Fake news here has led to major stints of violence among castes and religions, interfering with public policies.

Generally, in highly populated countries like India where the population is 1.2 Billion it is widespread because of the Internet and Smartphones. One such tool or application which facilitate this kind of widespread is WhatsApp. It is estimated that around 200 Million active users log in everyday and use this application to send messages.

Imagine how fast the news will traverse, and it is proved in many instances that false news or derogatory news moves faster than good and true news.

In the end of 2016, India launched a new currency in denominations of 2000, 1000 and 500-rupee notes. During this time there was a fake news which got released on internet and WhatsApp that there is a chip which is embedded in the rupee notes which will have some spying mechanism and can track them 120 meters underneath earth. Later ministry of finance had to plunge in and clear this misleading news. [5].

In this thesis we try and investigate whether machine learning methods can be of help to detect fake news for sure. Experts affirm that cooked up stories about Hillary Clinton

were one of the many reasons on the results she has secured such low results against Donald Trump in the presidential elections of USA.

Also, many governments like Finland are trying to address these problems at an early stage through education. All in all, it is very evident that the development of machine learning techniques for detection of fake news is a dire need!

1.1 Motivation

Fake News can be described as completely deceitful or cooked up information that is being broadcasted claiming as true facts. Identification of this false information in our daily life although is very much related to deception detection, however in true sense its much more difficult and complex.

Many societies in this world believe fake news and carry their thoughts blindly by believing in fake news. And some age groups do take wrong actions because of fake news which is of a greater concern to this mankind. In order to tackle this we have thought that machine learning would be of help for solving this problem.

Crafting an efficient and engineered solution for a classical supervised machine learning program to identify the fakeness is also a technically challenging task. It is hard for multiple reasons and the foremost reason being , manual task of identifying fake news is very subjective. Assessment of the exactitude of a news story is a complex and tiresome task, even for fully trained experts. TodayâĂŹs world source of news is through multiple channels like traditional media outlets, social media channels, user live feeds etc.

In this scenario of enormous amounts of fake news information, identification of the fakeness will be really helpful for the betterment of world in many ways.When we look deeper into this problem of fake news there is two broad motivational areas which needs high focus, one is detection of fake news and the second is the classification of fake news and its widespread areas and speeds in which it traverses.

Since this is a vast array of issues comprising of various sources like social media, internet, smart phone technology, connectivity, political rallies etc we wanted to contribute in an program which looks deep into the problem and finds the best fit process using machine learning and deep learning.

Fully automated solutions need deep understanding of the natural language processing which is not so easy in all means. These complexities make it a frightening task to classify text as fake news. In this thesis, we develop a deep learning based model for detecting fake news. Deep learning has the advantage in the sense that it does not require any

handcrafting of rules and/or features, rather it identifies the best feature set on its own for any specific problem.

1.2 Problem Definition

Consider a scenario in which we are given,

classification of news article given the statement of a news article A and the statement of news article B, task is to classify B into one of the three categories.

agreed: B debate or hold about the same fake news as A

disagreed: B contradict/disprove the fake news in A

unrelated: B is totally unrelated to A

The goal of this work is to efficiently determine the relationship between statements (whether the statement B can be reasonably inferred from statement A) and classify the relation as agreed,disagreed,unrelated.

1.3 Usecases/Examples

Image for Doc.png Image for Doc.png

	id	tid1	tid2	title1_en	title2_en	label
0	0	0	1	There are two new old-age insurance benefits for old people in rural areas. Have you got them?	Police disprove "bird's nest congress each person gets 50,000 yuan" still old people insist on going to beijing "	Unrelated
1	3	2	3	If you do not come to Shenzhen, sooner or later your son will also come. In less than 10 years, Shenzhen per capita GDP will exceed Hong Kong.	Shenzhen's GDP outstrips Hong Kong? Shenzhen Statistics Bureau dismisses rumors: only the gap is narrowing	Unrelated
2	1	2	4	If you do not come to Shenzhen, sooner or later your son will also come. In less than 10 years, Shenzhen per capita GDP will exceed Hong Kong.	The GDP overtopped Hong Kong? Shenzhen clarified: a little bit more	Unrelated
3	2	2	5	If you do not come to Shenzhen, sooner or later your son will also come. In less than 10 years, Shenzhen per capita GDP will exceed Hong Kong.	Shenzhen's GDP topped Hong Kong last year? Shenzhen Bureau of Statistics refutes rumors: 61.1 billion	Unrelated
4	9	6	7	How to discriminate oil from gutter oil by means of garlic	It took 30 years of cooking oil to know that one piece of garlic is easy to spot	agreed
5	4	2	8	If you do not come to Shenzhen, sooner or later your son will also come. In less than 10 years, Shenzhen per capita GDP will exceed Hong Kong.	Shenzhen's GDP overtakes Hong Kong? Bureau of Statistics refutes rumor: Unsurpass but the gap shrinks again	Unrelated
6	6	9	10	if you eat durian, you will kill yourself if you eat it wrongly!	Durian can't eat with anything, it's the same as coffee, it's heart disease.	Unrelated
7	5	2	11	If you do not come to Shenzhen, sooner or later your son will also come. In less than 10 years, Shenzhen per capita GDP will exceed Hong Kong.	.Shenzhen's GDP outpaces Hong Kong? Defending Rumors: The gap has narrowed yet again	Unrelated
8	7	12	13	Frog frog? It's a fertility test! Let's play" Jewel V ".	A store in xianning contains cotton"? A multi-agency association in chongyang university	Unrelated

Figure 1.1: sample dataset

sample Dataset from ByteDance .

1.4 Challenges

1.4.1 General Challenges:

Challenges of identifying fake news arises from the fact that it is even difficult for human beings to isolate fake news from original news.

Another crucial challenge is the use of language in fake news which makes it even more strenuous to identify fake news.

Typically we can classify few linguistic factors which constitute the creation of fake news such as subjective, intensifying and fudge words with an ambition to induce vague, obscuring, dramatizing or sensationalizing language.

In middle of all these complex factors , applying feature-based approaches will be labor-intensive and time-consuming.

Fake news generally has a combination of true stories with false details, which are baffled to read and understand the truth or fact. Most of the times the fake news creator merges true stories with false details with an ambition to mislead people.

For example, the statement However, it took 19.5 million dollar in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific is half-true since it combines the true number of 19.5 million dollar and the misleading place where the money went to. In such case, it is easy to get people's attention about trusted parts without noticing the presence of fabricated ones.

Fake news dataset for research and exploration is currently limited. Currently, there is only dataset for political fake news published for reaserch . Domains other than politics like education,bussiness,medical related datas are still not open to future research and exploration.

1.4.2 Challenges in dataset

Imbalanced dataset posses a different challenge in classification. The change in distribution percentage of the labels in the dataset makes many machine learning algorithms less efficient, mainly in the prediction of minority class.

1.5 Contributions

Most of the existing studies on fake news detection are based on classical supervised model. In recent times there has been an interest towards developing deep learning based fake news detection system, but these are mostly concerned with binary classification.

In this thesis, we attempt to develop an two way representation based deep neural network architecture for fake news detection. The individual models are based on Decomposable attention and Long Short Term Memory (LSTM). The representations obtained from these two models are fed for predicting the news statement pairs.

1.6 Outline

The remainder of the thesis is structured as follows. In Chapter 2, we discuss related work done and some important ideas used in the thesis. In Chapter 3 we provide a primer on various techniques used in our experiments such as natural language and deep learning and We also present the baseline model and we also explain the proposed solution to solve the fake news detection problem.

In Chapter 4, we present the experimental setup and results to evaluate our approach. In Chapter 5 we discuss and analyze the methods investigated .Finally conclusions and future work are presented in Chapter 6

Chapter 2

Related Work

The rise of fake news in recent years and the effects of it on the 2016 US elections, several studies have been conducted regarding fake news and fake news detection. In this section, we try to mention and analyze the important researches which we find relatable to our work. We categorize them into three subsections:

2.1 Text classification problem

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang [6] have proposed features such as nubmer of words, number of characters per word, frequencies of words and phrases ,parts of speech tagging (i.e., n-grams and bag-of-words approaches), partsof-speech (POS) tagging for the classification task

Shlok Gilda have proposed a solution for detecting fake news that suggest bi-gram TF-IDF gives effective models [7].

Many work suggested sentiment analysis for deception detection as some correlation might be found between the sentiment of the news article . Conroy, Rubin, Chen, and Cornwell hypothesized expanding the possibilities of word-level analysis [8].

Mathieu Cliche in his sarcasm detection blog has described the detection of sarcasm on twitter through the use of n-grams, . Wang has compared the performance of SVM, LR, Bi-LSTM, CNN models on their proposed dataset LIAR [26].[9]

Ferreira and Vlachos approached the stance classification task supervised by the Emergent dataset using a logistic regression classifier [10]

Linguistic features, such as lexical and syntactic features, capture specific writing styles and sensational headlines that commonly occur in fake news contents (Potthast et al.

2017), while visual features are used to identify fake images that are intentionally created or to capture specific characteristics for images in fake news (Gupta et al. 2013).[11]

knowledgebased: using external sources to check the authenticity of claims in news contents (Magdy and Wanas 2010; Wu etal. 2014), and (2) style-based: capturing the manipulation in writing style, such as deception (Rubin and Lukoianova2015) and non-objectivity (Potthast et al. 2017).[11]

2.2 Deep Learning for Fake News Detection

A.Arjun Roy, Kingshuk Basak, Asif Ekbal, Pushpak Bhattacharyya proposed an ensemble architecture based on CNN [12]and Bi-LSTM [3], and this has been evaluated on Liar dataset. model tries to capture the pattern of information from the short statements and learn the characteristic behavior of the source speaker

Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. Neural computation [3].

B.Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung [13] was proved to yield better results than previous feature-based learning they used Recurrent Neural Network (RNN) for predicting weather the tweets were rumors or not.they structured tweets as sequential data

Many other complicated approaches have also been investigated: Sahil Chopra and Saachi Jain. 2017 bidirectional LSTM/GRU architectures some with modifications [14].

ensemble of classifiers James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism [15].

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi proved better result of lstm through analysis of features [16].

2.3 Attention Models for NLP:

The main idea of the attention mechanism in natural language processing is soft selection of sequence of words based on their importance of task. alignment, attention (Bahdanau et al., 2015) [17]

pairwise attention to model query-document pairs.(Xiong et al., 2016; Seo et al., 2016) [18] Dynamic coattention networks for question answering.

Attention been predominantly used in conjunction with LSTMs(Rocktaschel et al., 2016; Wang and Jiang, 2016)[19]

attention to a lesser extent with CNNs (Yin et al., 2016).[20]

Our method is motivated by alignment, decomposable attention(The insight is that it can be enough to align (contradicting) words pair-wise and then aggregating that information) and they are largely independent of word order A Decomposable Attention Model for Natural Language Inference Ankur P. Parikh [21].

Intuitively, the relationships between two words is often not straightforward. Words are complex and often hold more than one meanings (or word senses). As such, it might be beneficial to model two way representations.we have modeled our solution based on decomposable attention and on sequential composition (LSTM).

Both views are used for prediction. To the better predictive performance, our final method employs an ensemble of these two representation.

Chapter 3

Solution Approach

3.1 Introduction

In this chapter we will discuss the relevant machine learning methods used in this thesis that form the basis of the solution. The chapter is structured as follows,First we describe the concept and methods of machine learning ,secondly the ways of representing text documents . Lastly we cover various techniques for training a model as well as tackling the problem of training a classifier on an imbalanced dataset.

3.2 Theoritical Background

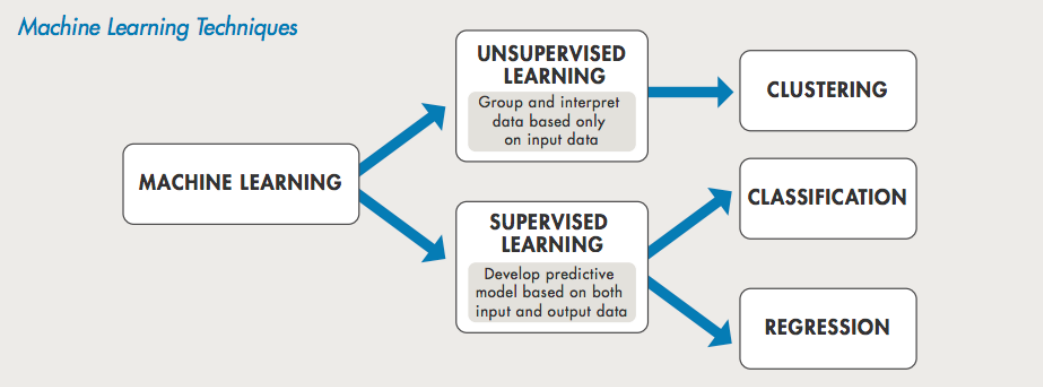

Figure 3.1: Mchine Learning source:Google

3.2.1 Machine Learning

Machine Learning is a approach which lets on the machine to study from examples and past informations , without particulary being programmed,i.e data is been feed to the machine by generic algorithm, and the algorithm builds the logic based on the provided data and predits the output.

To create a model a training dataset is provided to train the Machine Learning algorithm,and when the input data is given to the algorithm the prediction is made on the basis of the model created.

The prediction is assessed for accuracy and if it is accepted, the Machine Learning algorithm is installed.

If the accuracy is not approved , the algorithm is trained again and again with the training data set. machine learning models need a lot of training data in order for the model to perform well. [22].

Machine learning algorithms are used in a wide variety of applications, such as email filtering,fraud detection ,online recommendation,automated inspection,face detection

Machine learning is sub-categorized to three types

UnSupervised Learning

It is a is machine learning task where the model learns through observation and finds arrangement in the data. Once a dataset is given to the model, it identifies pattern and relations by developing clusters in it but to the cluster label can not be attacted .

for example in a group group of apples or mangoes, it will separate all the apples from mangoes.if a new data is given to the model based on the pattern it will group the data Supervised Learning: [22].

Supervised learning

is machine learning task where the model is getting trained on a labelled dataset. The algorithm analyses the data and gets inferred later uses for mapping new data

i.e function that maps an input to an output. once the training is done its start making the prediction on the new data provided. [23].

In Supervised learning let the input variables (x) and an output variable (Y) and algorithm is used to learn the mapping function f(X) from input to the output.

$Y = f(X)$

The algorithm examines the training data and reproduce the approximate function, which can be termed mapping function. This function can then be used for predicting new input dataset.

The algorithm will perfectly predict unseen dataset. Thus supervised learning algorithm uses statistical ways to map function from training dataset to unseen dataset.

Supervised learning are classified into regression and classification problems.

1.Classification:

It predicts the a category of the data to which they belongs to. Model is trained to predict the qualitative outputs. eg: Spam Detection, Sentiment Analysis.

2.Regression:

In regression problem one needs to predict a output variable, output value is an integer or float.

3.2.2 Deep Learning

Deep learning is a subset of machine learning that uses multi layer neural networks to transfer data from input to output and to deliver higer accuracy in tasks such as detection, speech recognition, translation .

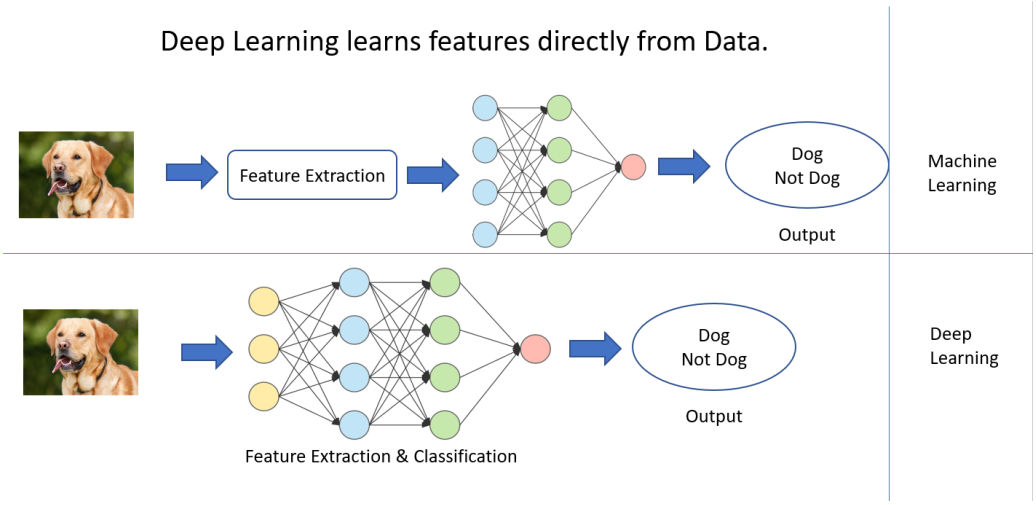

Figure 3.2: neural network:source :Google

Deep learning are different from traditional machine learning techniques , they can automatically learn features from data such as text, video or images, without human interference or implicitly given.

Neural network architectures directly learn from raw data and can improve their prediction accuracy when data is more provided.

Deep learning is responsible for many new happenings like self-driving cars, voice assistants .

Neural Network

The deep learning fundation are the neural network. neural network represent a structure, that combines machine learning algorithms for solving many specific tasks.

Neural Networks consist of layers,each layer has many neurons. A connection between these neurons is compared to the neurons in brain which hold the synapse . where the signal are transfered from one neuron to another.

Neural network is comprised of several layers, each layer perfors a specific function that posses way for solving the problem .

The first layer is the input layer, where the input is fed.for example if speech recognition task then the input would be the speech data.

Next to the input layer is one or more hidden layers, followed with an output layer. The output layer has the final solution to the problem,Real learning takes place at the hidden layers. [24].

The data is transferred , from the input, to the hidden layers, and ends in the output layer. This is refered feedforward neural network.

deep learning systems has many neural networks that are trained using huge amount of data

RNN: Recurrent Neural Networks

The RNN architecture is not dissimilar to the neural network. The only difference is that the neural network has a new concept of memory, and it exists in form of link.

the outputs of the layers are fed back into previous layer as inputs is used for analysis of sequential data, which is nor in neural network.

RNN has no fixed-length input, such restriction is in traditional neural network.

The links between layers in the reverse direction is feedback loops, they help to learn the concepts based on context.

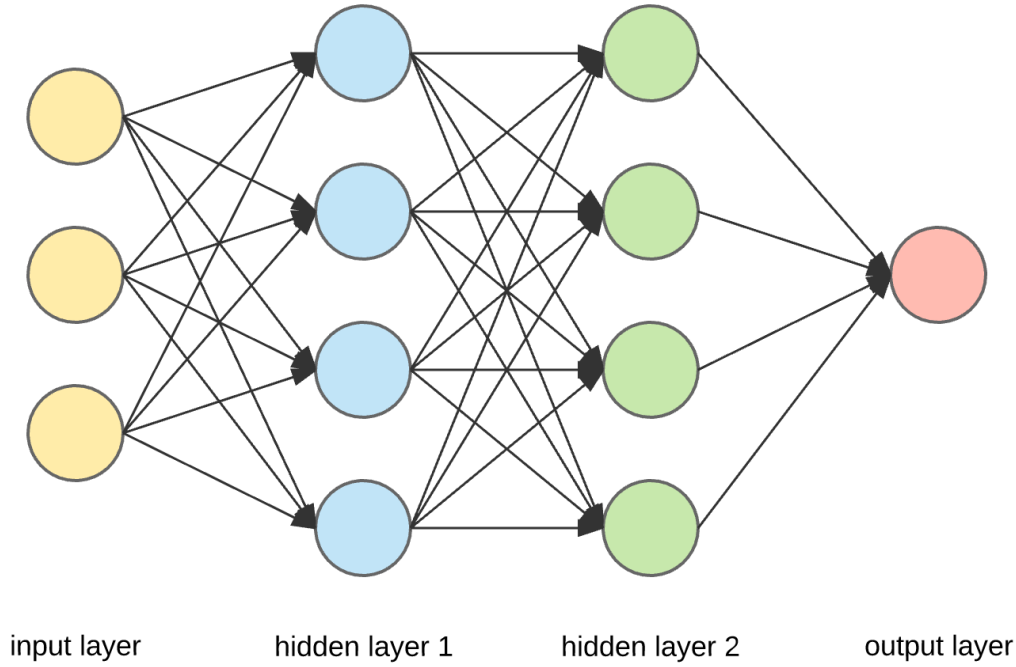

Figure 3.3: neural network:Google

The RNN is able to recognize time series data and take advantage of the time-related context that contains repeated patterns. This is done by adding weight to patterns where the tokens are recognized,

framework for learning sequence data is called the Long Short-Term Memory Network (LSTM). It is a type of RNN that is capable of learning long-term relationships. LSTM is designed to assist with dependent on context. [24].

Rnn popular example:

Image classification, Image captioning,

Convolutional Neural Networks

it performs the convolution operation in certain layers hence, the name Convolutional Neural Network. The architecture changes from the traditional NN,

In CNNs, the first layer is always a Convolutional layer. with dimensions: length, width, and depth.

They are not fully connected I.e the neurons from one layer do not connect to each and every neuron in the following layer. The output of the final layer is the input to the first fully connected layer.

In other words convolution is grouping function that takes place between two matrices.CNN merges information. In practice it is like feature selection.

it is common to add pooling layers in between convolution layers. It is responsiblefor reducing its dimensionality. Thus training time is reduced , and the problem of overfitting is also solved. Next come the fully connected layers.

Finally fully connected layers, has features most accurately for all classes , and the final output is the sigle neuron. For example, vehicle recognition

A side-view picture of a car has two wheels. incomplete description will match a motorcycle. additional features such as the presence of windows and/or doors will help to more accurately determine the vehicle type.

some deep learning applications may benefit from the combination of RNN and CNN [24].

3.2.3 Representation of text

The bag-of-words model

represents statements,sentences as multiset of words, all occurrences of an element is stored but order is ignored. I.e any spatial information is not captured in bag of words model, such as wordword co-occurrence .

in n-gram spatial information is captured by storing the occurrences of n words appearing in sequence in the document.a text is represented as set words, with no grammar and word order but keeping multiplicity.

The bag-of-words model is commonly used in methods of document classification .

In this model, a text is represented as set words, with no grammar and word order but keeping multiplicity. The bag-of-words model is commonly used in methods of document classification [25]

TFIDF

tfidf term frequency-inverse document frequency. Tf-idf weight is intended to show how important a word is for sentence or document in a collection or corpus

Tf-idf stands for term frequency-inverse document frequency, and it is used in information retrieval and text mining. This weight is how important a word is to a document in a collection or corpus.

The importance increases to the number of times a word appears in the document . Variations of the tf-idf weighting scheme are central tool in scoring and ranking a document's relevance when a given a query.

Tf-idf used for stop-words filtering in text summarization and classification. [26].

the tf-idf weight is composed by two terms:

TF: Term Frequency: measures how frequently a term occurs in a document. document are of different length, a term would appear many times in long documents than shorter ones.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, how important a term is. Weighing down the frequent terms and scale up the rare ones

IDF(t) = loge(Total number of documents / Number of documents with term t in it).

one-hot-vector

is a vector where a single element is one and the rest of the elements are zero. The index of the element set to one in the one-hot vector maps to the word in the vocabulary at that index. However, this means that the dimension of the one-hot representation increases with the size of the vocabulary used. Thus, in a large corpus the one-hot encoding of words tend to become an excessively large and sparse representation.

One of the major problems with Machine Learning is the fact that you cannot work directly with categorical data. Machine Learning cannot work with categorical data.it is bunch of mathematical operations translated to a machine understandable.

So input data is converted in to numbers.. Once we are done with the processing on the numbers, we need another mechanism to somehow revert the output to the same format as that of the input data. This is where One-Hot Encoding does in.

Word2vec: is a group of models utilizing either the continuous bag-of-words model architecture or the continuous skip-gram model architecture, where the former predicts the current word given the context window of surrounding words while the latter predicts the context window of surrounding words given the current word

Word2vec is efficient for learning word embeddings from raw text. Two types Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. CBOW-target words is predicted from source context words,skip-gram predicts source context-words from the target words

Glove

Global Vectors, is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

It is developed as an open-source project at Stanford. [27] As log-bilinear regression model for unsupervised learning of word representations, it combines the features of two model families, namely the global matrix factorization and local context window methods In this paper, we experiment the model with Glove word embedding representations

GloVe -global vectors for word representation. It is developed by Stanford for generating word embeddings by aggregating global word-word co-occurrence matrix from a corpus. The resulting embeddings show substructures of the word in vector space

3.2.4 Training

Training a neural network with given a labelled dataset using supervised learning intent to find the network parameters for minimizing the error rate . The procedure derives a function that links a given input to a class label. The target is that the learned function can be used for mapping unseen inputs as well, called generalization, which requires the function to model the fundamental relationship in the labelled examples from the training dataset.

Loss Function

The loss function is a function which is used to measure difference between the actual labels and the predicted output labels. It is a non-negative value, where the stability of model increases along with the decrease of the value of loss function. A commonly used loss function for classification tasks during trainng is the cross-entropy loss function, which measures how close the output probability distribution of a classifier [28]

Squared Error

Mean Squared Error (MSE) is basic of loss functions: itâĂŹs easy to implement and works well. T MSE is calculated as, difference between predictions and the ground truth, square it, and average it across the dataset. [28]

Likelihood Loss

The likelihood function is used in classification problems. The function takes the predicted probability for input and multiplies . it is useful for comparing models.

For example, model probabilities of [0.4, 0.6, 0.9, 0.1] for the ground truth labels of [0, 1, 1, 0]. The likelihood loss calculated as (0.6) * (0.6) * (0.9) * (0.9) = 0.2916. we multiply the modelâĂŹs outputted probabilities together for the actual [28] outcomes. Log Loss (Cross Entropy Loss)

Log Loss is used in classification problems, Kaggle competitions most popular measures. straightforward modification of the likelihood . it penalizes very confident and very wrong. Predicting the wrong class with high probablity. [28]

Gradient Descent

Gradient descent is an iterative optimization algorithm for finding minimum of a function by taking small steps in the direction of the negative gradient [29] . It is called a first-order optimization algorithm because the first derivative is taken into account when performing the updates on the parameters.

The algorithm iterates over dataset guiding the parameters with the aim of converging where the network minimizes the loss function. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate Îś. Therefore, we follow the direction of the slope downhill until we reach a local minimum The algorithm needs several epochs training a neural network before reaching the convergence criteria.

For a feedforward neural network stacking multiple layers on top of each other, the gradients of parameters in the hidden layers are derived applying the chain rule for each layer. In order to calculate the error contributions of the parameters, the loss calculated at the output of the network must be propagated back through the layers.

This technique is also called backpropagation or backward In a deep neural network forward propagation through every layer inorder to get the predicted value. the loss is calculated at the output of the network by propagating back through the layers. This technique is also called backpropagation

Generalization

There is a terminology used in machine learning when we talk about how well a machine learning model learns and generalizes to new data, namely overfitting and underfitting.

poor performance of machine learning algorithms are mainly due to Overfitting and underfitting. A model that overfits on the training data shows poor predictive performance on unseen examples.when a model is too complex or iterative training procedure is performed too long overfitting occurs.

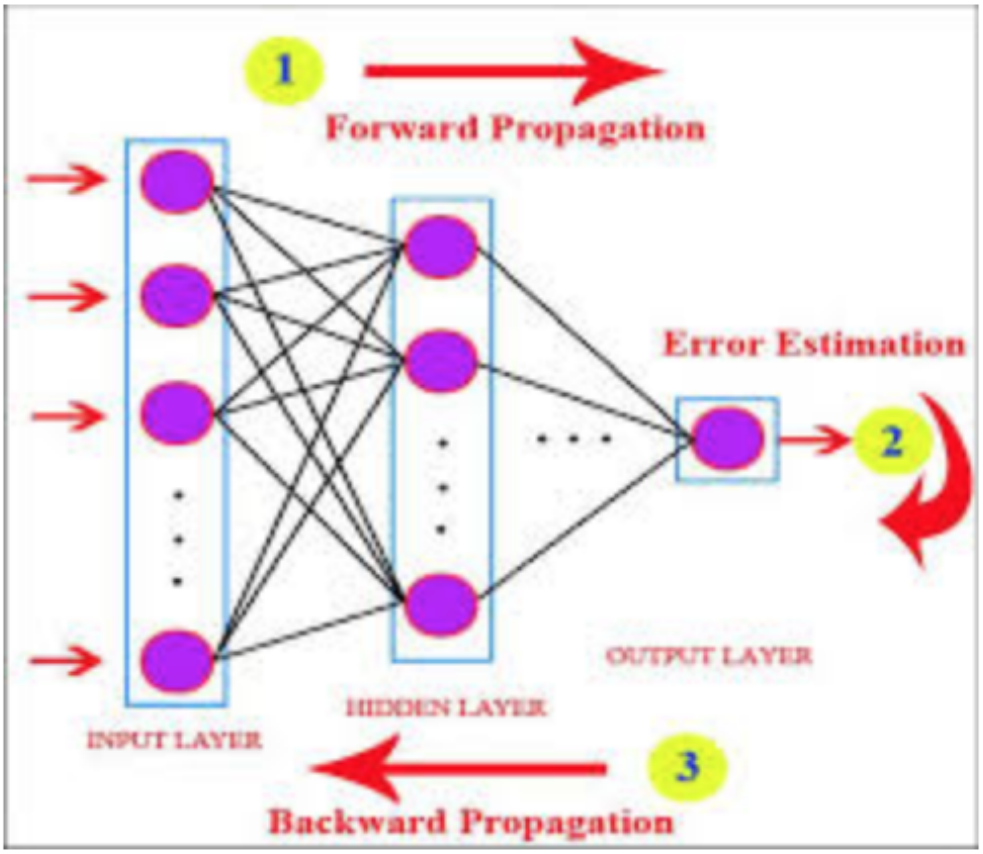

Figure 3.4: Confution Matrix source:Google

In supervised learning, one speaks of overfitting when the function inferred tends to describe outliers and noise instead of capturing the underlying relationship in the training data.

A model that overfits on the training data shows poor predictive performance on unseen examples, implying a generalization error. Overfitting occurs when a model is too complex. For neural networks the problem of overfitting commonly occurs when there are too many parameters relative to the number of training examples and/or when the iterative training procedure is performed too long.

Early stopping is commonly used when training a neural network to prevent the model from overfitting on the training dataset. This technique stops the iterative training procedure when the performance on unseen examples starts to decrease. Early stopping is a form of regularization .Regularization is aimimg to improve gereraliaton by induce less complex model.

When training neural networks, large weights are penalized by addind L1 and/or L2 regularization to the loss function . The L1 penalty term is the sum of the absolute value of all weights and the L2 penalty term is the sum of the square of all weights. Dropout is another commonly used technique where randomly selected neurons are ignored during training to Prevent Neural Networks from Overfitting [30].

Class Imbalance

In a dataset Imbalanced data typically refers to the classes in classification problems which are not represented equally . The skewed distribution makes many conventional machine learning algorithms less effective, especially in predicting minority class.

Performance Metrics

With imbalanced classes, without actually making useful predictions it is easy to get a high accuracy.if the class labels are uniformly distributed then accuracy as an evaluation metrics makes sense . So regarding imbalanced classes confusion-matrix is good technique to summarizing the performance of a classification algorithm.

	positive prediction	negative prediction
positive class	true positives (TP)	false negatives (FN)
negative class	false positives (FP)	true negatives (TN)

Figure 3.5: Back propagation source:Google

true positives (TP) denoting the number of positive instances that a classifier correctly predicted as positives, false negatives (FN) denoting the number of positive instances that a classifier incorrectly predicted as negative

Derived from the results in a confusion matrix, the F1-score,precision,recall calculated and widely used as a Performance Metrics in imbalanced domains [31].

True Positives (TP) This is the count of rightly predicted positive values.

True Negatives (TN) - rightly predicted negative values.

False Positives (FP) wrongly predicted positive value which real class is no and prediction is made yes class.

False Negatives (FN) wrongly predicted negative real class is yes but prediction is made in class .

Accuracy

classification model accuracy is a performance measure and it is ratio of rightly predicted values to the total values. higher the accuracy, better the model.

This is correct when the dataset is almost symmetrical where FP and FN are nearly same . to evaluate the performance of the model other measure should also be considered.

implies the model is approx. 86 percentage accurate when the learning algorithm used is two way representation. Accuracy = TP+TN /TP+FP+FN+TN [32].

Precision

rightly predicted positive values to the total positive predictions. Higher the value of precision, lower is the false positive rate. In this study precision of 0.85 has been achieved with our proposed solution.

Precision = TP/TP+FP [38]

precision and recall both responsible for accuracy of the model. Precision is results percentage which are relevant. recall is total relevant resultsre classified by algorithm. [32]

Recall

rightly predicted positive values to the all value,sensitivity model.predicted corectly

Recall = TP/TP+FN [38]

For example, for a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned.

In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

F1 score

F1- false positives and false negatives into account and weighted average of recall and precision. F1 is useful than accuracy, data distribution is uneven. making wrong prediction cost is different,

F1 Score = 2 (Recall$\times Precision)/(Recall + Precision)$

F1 score interpretation:

F1 score is high, both precision and recall of the classifier shows good results.

the metric allows us to compare the performance of two classifiers using just one metric F1.

If F1 score is low, we can not judge the false positives or false negatives. best way is to use confusion matrix to diagnose the problem and then look at validation or test dataset. [32]

Balancing the Class Distribution

Mainly there are two techniques for balancing the class distribution of a dataset:

(i) over-sampling - minority classes randomly replicating examples- risk of overfitting and

(ii) under-sampling - majority classes randomly removing examples- lose important information [33].

3.3 Existing Approaches/Baselines

Modeling approaches:

In this chapter we describe the models we investigate for representing the relationship between the news statement pair. Here we present how these models are used for the task

3.3.1 XGBoost

XGBoost stands for eXtreme Gradient Boosting. [34]

XGBoost has become a widely and popular tool Kaggle competitors and Data Scientists in industry. It is a highly versatile and flexible so most regression, classification and ranking problems are been solved by xgboost.

gradient boosting is an ensemble method that corrects previous models by sequentially adds predictors . However, after every iteration instead of assigning different weights to the classifiers , using the previous prediction this method fits the new model and then minimizes the loss when adding the latest prediction.

So, finally using gradient descent the model is updated and hence XGBoost is called gradient boosting.It is used for both regression and classification problems.

XGBoost ensemble learning method. rely upon just machine learning model for result is not often sufficient. Ensemble learning combine the predictive power of multiple learners.

Bagging and boosting are two widely used ensemble learners.

Bagging

decision trees are interpretable models, . training dataset that we split into two parts. to obtain two models each part to train a decision tree.

When we fit both these models, they would produce different results. Decision trees has high variance because of behavior. Bagging or boosting helps to reduce the variance . . The final prediction is the averaged output .

Boosting

In boosting, subsequent tree goal is to reduce the errors of the previous tree. tree learns from its predecessors and residual errors is updated.

weak learners are base learners in boosting . weak learners contributes some vital information for prediction, The final strong learner brings down both the bias and the variance.

boosting makes use of trees with fewer splits rather than bagging techniques in which trees are grown to their maximum extent. large number of trees lead to overfitting. So, stopping criteria is necessary for boosting.

Boosting consists of three simple steps:

F0 is defined to predict the target variable y. This model will be associated with a residual (y-F0)

h1 is fit to the residuals from the previous step

Now, F0 and h1 are combined to give F1, the boosted version of F0. The mean squared error from F1 will be lower than that from F0:

To improve the performance of F1, we could model after the residuals of F1 and create a new model F2:

This can be done for âĂŸmâĂŹ iterations, until residuals have been minimized as much as possible:

Here, the additive learners do not disturb the functions created in the previous steps. Instead, they impart information of their own to bring down the errors.

3.3.2 LSTM :Long short term memory

Recurrent Neural Networks suffer from short-term memory so it is hard to carry information from beginning time steps to end step. So it is possible for leaving out some important information from the early time step.

During back propagation, RNN suffer from the vanishing gradient problem. Gradients are values used to update a neural networks weights.

If the gradients tend to vanish it will be difficult for the RNN to capture long-term dependencies in sequences and, conversely, if the gradients tend to explode it might be difficult training the network due to changing weights.TO overcome this short term memory LSTM were introduced by Hochreiter and Schmidhube [3].

The idea of an LSTM unit is that it has a internal mechanisms called gates that regulates how much of the information is added to this and how much to forget . From this information it passes the relevant data down the long chain of sequences to make predictions

Our LSTM model was pre-trained with 100-dimensional GloVe embeddings. Output dimension and time steps were set to 300. ADAM optimizer with learning rate 0.0001 was applied to minimize categorical cross entropy loss and Relu was the activation function for the final output layer. Finally, this model was trained over 10 epochs.

$$
\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}^i\mathbf{x}_t + \mathbf{U}^i\mathbf{h}_{t-1} + \mathbf{b}^i\right), \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}^f\mathbf{x}_t + \mathbf{U}^f\mathbf{h}_{t-1} + \mathbf{b}^f\right), \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}^o\mathbf{x}_t + \mathbf{U}^o\mathbf{h}_{t-1} + \mathbf{b}^o\right), \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh\left(\mathbf{W}^c\mathbf{x}_t + \mathbf{U}^c\mathbf{h}_{t-1} + \mathbf{b}^c\right), \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right),
\end{aligned}
$$

Figure 3.6: lstm equation

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network sequence prediction problems in network is capable of learning order dependence.

This is a behavior required in complex problem domains like machine translation, speech recognition, and more.

LSTMs are a complex area of deep learning like bidirectional and sequence-to-sequence relate to the field.

3.4 Proposed Solution

Most of the existing studies on fake news detection are based on classical supervised model. In recent times there has been an interest towards developing deep learning based fake news detection system. In this thesis, we attempt to develop an two way representation based deep neural network architecture for fake news detection. The individual models are based on Decomposable attention and Long Short Term Memory (LSTM). The representations obtained from these two models are fed for predicting the relationship between the news statement pairs .

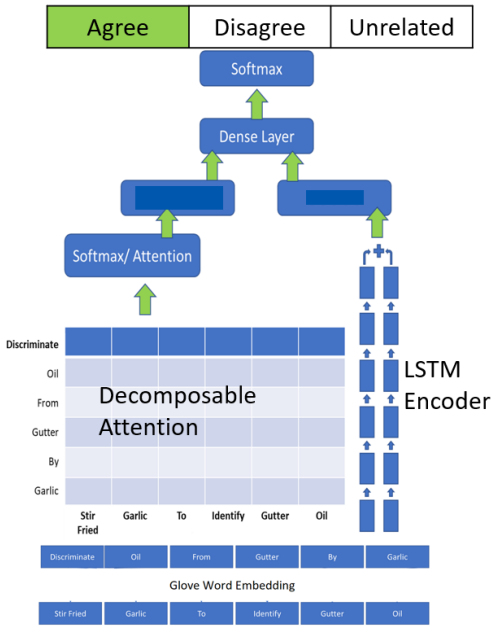

Figure 3.7: Proposed Solution

Intution

the intution behind this process is the relationships between two words is often not straightforward. Words are complex and often hold more than one meanings (or word senses). As such, it might be beneficial to model two way representations , one based on decomposable attention and another based on sequential composition (LSTM). Both views are used for prediction.

Novelty about this approach

Decomposable attention mechanism(Ankur P. Parikh)simply considers a word pair interaction and does not model the input document sequentially.it is beneficial to use a separate compositional encoder for this purpose, i.e., learning compositional representations we employ the standard Long Short-Term Memory (LSTM) encoder.So novelty about the solution is two way representation of the text is fed into the predictive layer for efficiently classify the relationship.

Decomposable attention

The insight is that deep modelling of the sentence structure is not required it is enough to align (contradicting) words pair-wise and then aggregating that information [35].

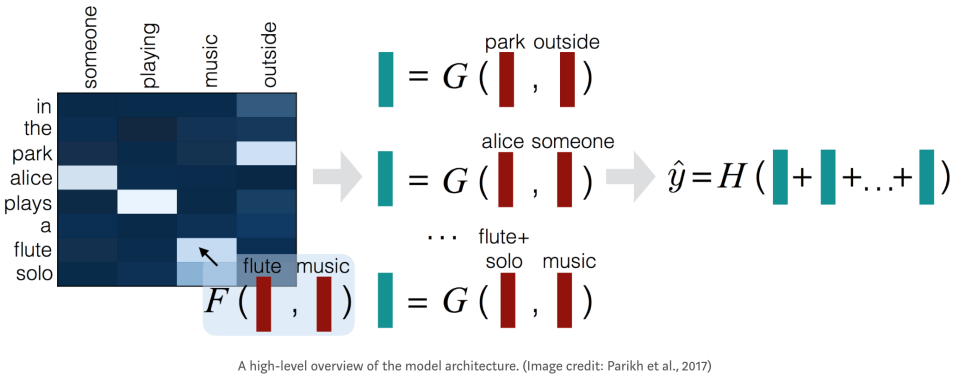

Figure 3.8: Decomposable Attention source:parikh Decomposable Attention

"thunder" and "lightning" aligned with "sunny" indicate a contradiction

Embed

All the words are mapped to their corresponding word vector representation.we used 100-dimensional GloVe embeddings

Attend

Given two statement a and b, For each word i in a and j in b, obtain unnormalized attention weights

$$e(i,j) = F(i)^T F(j)$$

where F is a feed-forward neural network. For i, compute a , $\beta$by performing softmax-like normalization of j using e(i, j) as the weight and normalizing for all words j in b. , βi captures the subphrase in b that is softly aligned to a. Similarly compute , αj for j. [36]

Compare

Create two set of comparison vectors, one for a and another for b For a, v1, i = G(concatenate(i, , β i)). Similarly for b, v2, j = G(concatenate(j, , αj)) G is another feed-forward neural network.

Aggregate

Aggregate over the two set of comparison vectors to obtain v1 and v2.

LSTM:

Clearly, decomposable attention mechanism(Ankur P. Parikh)simply considers a word pair interaction and does not model the input document sequentially.it is beneficial to use a separate compositional encoder for this purpose, i.e., learning compositional representations we employ the standard Long Short-Term Memory (LSTM) encoder. The output of an LSTM encoder at each time-step can be briefly defined as: $,hi = LSTM(w, i)$.LSTM is a specialform of recurrent neural networks (RNNs), which process sequence data

The inputs to the LSTM encoder are the word embeddings right after the input encoding layer and not the output of the intraattention layer.

The input text and each text is mapped with vector using glove 100 d word embedding.lstm layer on the embedding and the two sequence are concatenated and fed into dense layer.

Prediction layer

Feed the aggregated results,lstm encoder representationthrough the final classifier layer. [37]

categorical-entropy loss function with relu activation function and softmax is used to predict the label.

Chapter 4

Experimental Evaluation

4.1 Experimental Setup and Data Set

In this section, we provide details into the dataset and data preprocessing and experimental results .

Dataset overview

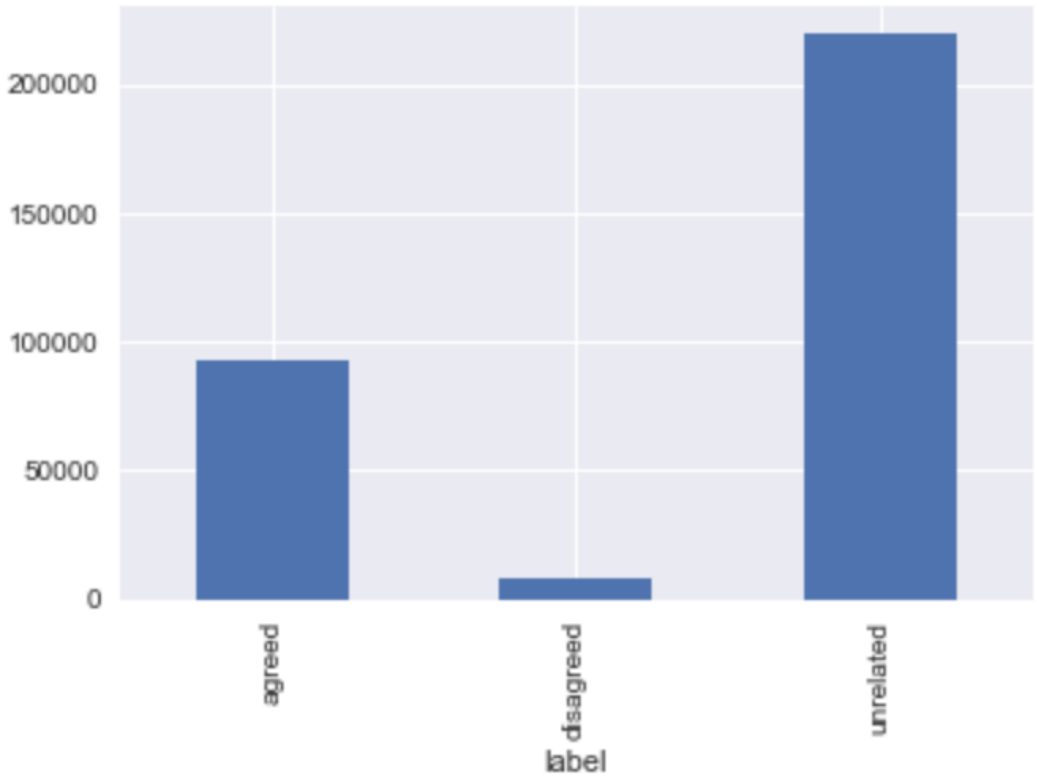

Figure 4.1: dataset label distribution

The datasets for our task are provided by ByteDance organization as . ByteDance is a global Internet technology company started from China . Dataset can be downloaded on kaggle page.

he complete training set contains 320,767 news pairs,these news pairs are labelled with relation as agreed/disagreed/unrelated.

The label distribution on the training dataset are unrelated 219313 agreed 92973 disagreed 8266

In order to build our classification models, we selected 80 percent of the data for training, 10percent of the data for validation and 10percent of the data for testing We also preprocessed our dataset extensively in order to split sentences, normalize casing, handle punctuation and other non-alphabetic symbols, and otherwise improve token consistency.

We also built vocabulary lists for terms occurring in the news statement pairs , and extracted corresponding pre-trained word embeddings for these tokens from the Stanford GloVe corpus (100d vectors from 6B corpus,).

Implementations tool

Keras

Keras is an open-source neural-network library written in Python.it is Designed to work with with deep neural networks which enable fast experimentation, it is user-friendly and extensible tool.

Keras models are (Sequential and Functional). Configure the layers

There are many tf.keras.layers available with parameters:

activation: Set the activation function for the layer. T By default, no activation is applied.

kernel initializer and bias initializer: creates the layer's weights (kernel and bias). This defaults to the "Glorot uniform" is default initializer.

kernel regularizer and bias regularizer: The regularization schemes that apply the layer's weights (kernel and bias), such as L1 or L2 regularization. compile class is called to build the model.

Scikit-learn

software machine learning library and it is freely available, coded in Python and Cython. It is very useful for modelling medium-scale supervised and unsupervised problems. .The

main advantages of using scikit learn is user friendly . Its performance, and good API consistency are also good . Scikit learn features a lot of classification, clustering and regression algorithm like gradient boosting. and is designed to work with libraries NumPy and SciPy.

NLTK:Natural Language Toolkit

NLTK is a Python package designed to help work with human language data. It has range of tools for processing language data e.g., tokenization, stemming,and part-of-speech tagging

Data Preprocessing

Text data requires preprocessing to implement machine learning or deep learning algorithms on them. There are many preprossing techniques widely used to convert text data into a form that is accepted for modeling. The data preprocessing steps that we used in our thesis

Part-of-Speech sentences might have excessive use of interjections as characteristic structures . Such structures might be captured by POS-tagging. Each tag is used as a feature.

Stop-words Removal

stop word are functional words do not carry any information but occur frequently like pronouns,propositions,conjunctions [36]. there are more than 400 stop- words if English is choosen as a language. to, the and are stopword example etc. when analyzing text they normally are of no use.

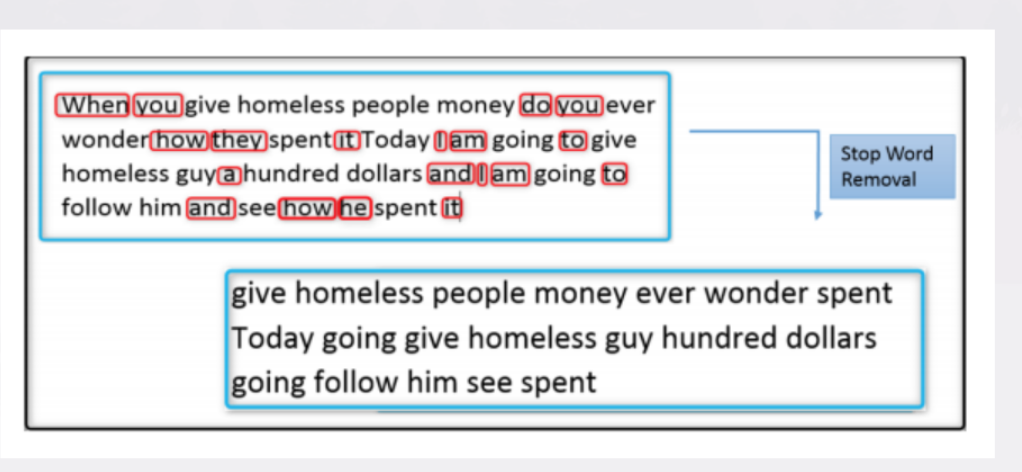

Figure 4.2: stop word removal source:google

This is done in scikit learn by passing (stop words=english).

Punctuation Removal

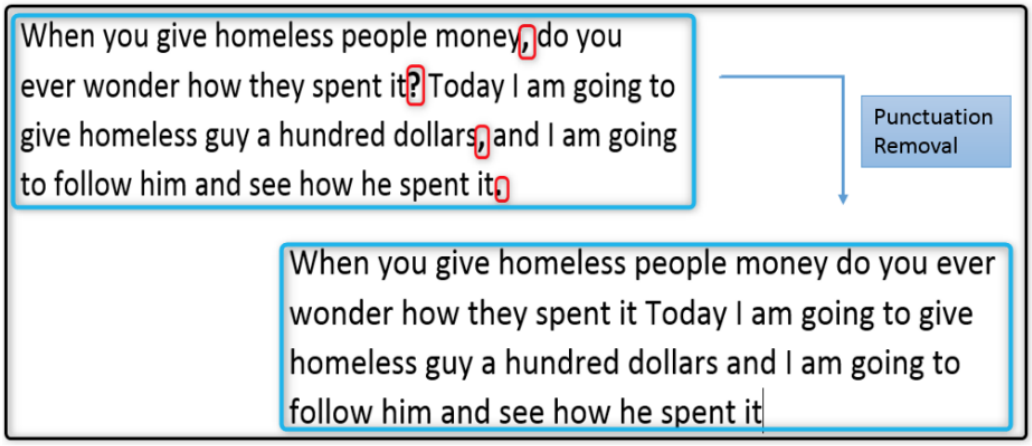

Figure 4.3: puntuation source:google

Punctuation is the use of spacing, conventional signs aids to the understanding and correct reading of written text .

Stemming Stemming is a technique to remove prefixes and suffixes from a word, ending up with the stem. Using stemming we can reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Figure 4 shows the example of stemming technique.

Stemming is a technique which identifies root of a word. For example, the words, attendinging, attentiotion, attends all can be stemmed to the word ATTEND. Stemming changes words to their roots word. stemming is done because root word typically describe same or close concepts in the document and therefore those words can be stemed . In this thesis , snowball stemming algorithm is used.

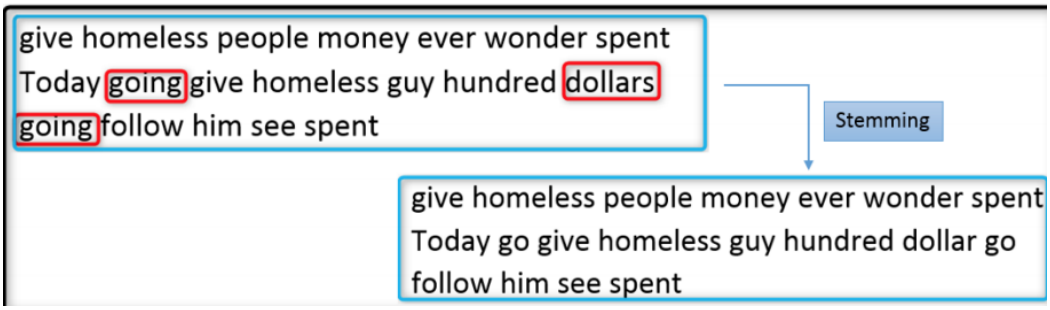

Figure 4.4: Stemming source:google

4.2 Experimental Results

In this chapter we describe the experiments of the thesis along with the results obtained. The models we present here were trained on the bytedance organisation dataset and for all experiments we report the performance of all model on the training and test dataset of bytedance

The models we implemented in the thesis:

XGBoost

XGBoost is open source library give high-performance implementation of gradient boosted decision trees.[wiki]

XGBoost's hyperparameters for model building:

learningrate: step size shrinkage used to prevent overfitting. Range is [0,1] -0.1

maxdepth: determines how deeply each tree is allowed to grow during any boosting round.we have choosed 50

subsample: percentage of samples used per tree. Low value can lead to underfitting.-0.8

colsamplebytree: percentage of features used per tree. High value can lead to overfitting. -.7

nestimators: number of trees you want to build. -80

objective: determines the loss function to be used like reg:linear for regression problems, reg:logistic for classification problems with only decision, binary:logistic for classification problems with probability. objective='multi:softmax'

alpha: L1 regularization on leaf weights. A large value leads to more regularization.-regalpha=4

Then from sklearn's modelselection using the the traintestsplit function we create the train and test set for cross-validation of the results with testsize size equal to 20 percentof the data.

Also, to maintain reproducibility of the results, a randomstate is also assigned.we instantiated XGBClassifier() class.

The following were the results we got and the accuracy we got in **xgboost.8294**

LSTM:Long short term memory

XGBoost	Precision	Recall	F1Score	support
Agreed	.69	.88	.69	30496
Disagreed	.76	.14	.24	2721
Unrelated	.90	.68	.78	72566
Avg/Total	.79	.76	.78	105783

Table 4.1: XGBoost REPORT

exploding and vanishing gradient problems were encountered when training traditional RNNs so LSTM LSTMs were developed to deal with them.

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.these cell remembers the follow of information. Words were represented using 100-dimensional GloVe word vector pretrained wordembedding.

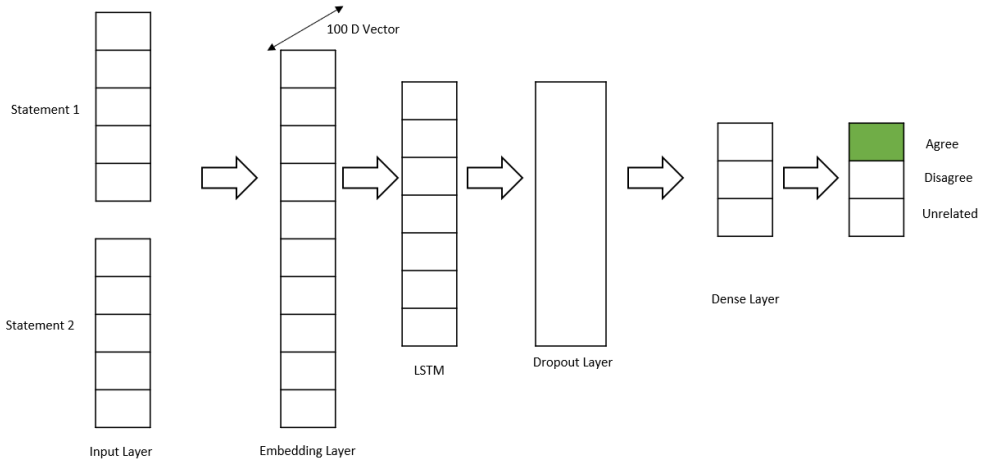

Figure 4.5: LSTM model

We padded sentences to equal length . The word vectors for the additional tokens were initialized with zeros. We preprocesed the data When training the classifier we minimized the cross entropy of the training set Adam optimization algorithm.The model architecture is presented in

LSTM	Precision	Recall	F1Score	support
Agreed	.84	.60	.69	30496
Disagreed	.85	.21	.34	2721
Unrelated	.83	.95	.88	72566
Avg/Total	.83	.83	.82	105783

Table 4.2: LSTM REPORT

Parameter Nam	Value
Activation Function	Softmax, ReLU
Dropout Rate	0.1
Epochs	50
Optimiser	Adam
Loss Function	categorical_crossentropy
L2 penalty	0.0001

Figure 4.6: LSTM model details

The following were the results we got and we had 100 epoch for accuracy **LSTM.8452**

The two way representation:The novelty approach.

This approach is motivated by AnkurPharik decomposable attention(focus on a subset of input data).The relationships between two words is often not straight forward. Words are complex and often hold more than one meanings (or wordsenses).

As such, it might be beneficial to model two way representations , one based decomposable attention and another based on sequential composition (LSTM). Both views are used for prediction.so the intution is having two ways of word order.

Decomposable attention:

Embed,attend,compare and aggregate are the four step involved in decomposable attention.we used glove embedding for mapping words we obtain unnormalied attention weights,by performing softmax obtain the subphrase of one sentence that is softly aligned to other.compare and aggregate.

Lstm The decomposable attention does not follows the word order.for learning compositional we employ the standard long short term memory. the inputs to the LSTM encoder are the word embeddings right after theinput encoding layer. The two representation is feed in the prediction layer.we used0.1. Dropout regularization (Srivastava et al., 2014)was used for all ReLU layers, and for the final linear layer we chose dropout for.02 and we obtain the result for our **proposed solution.8593** .

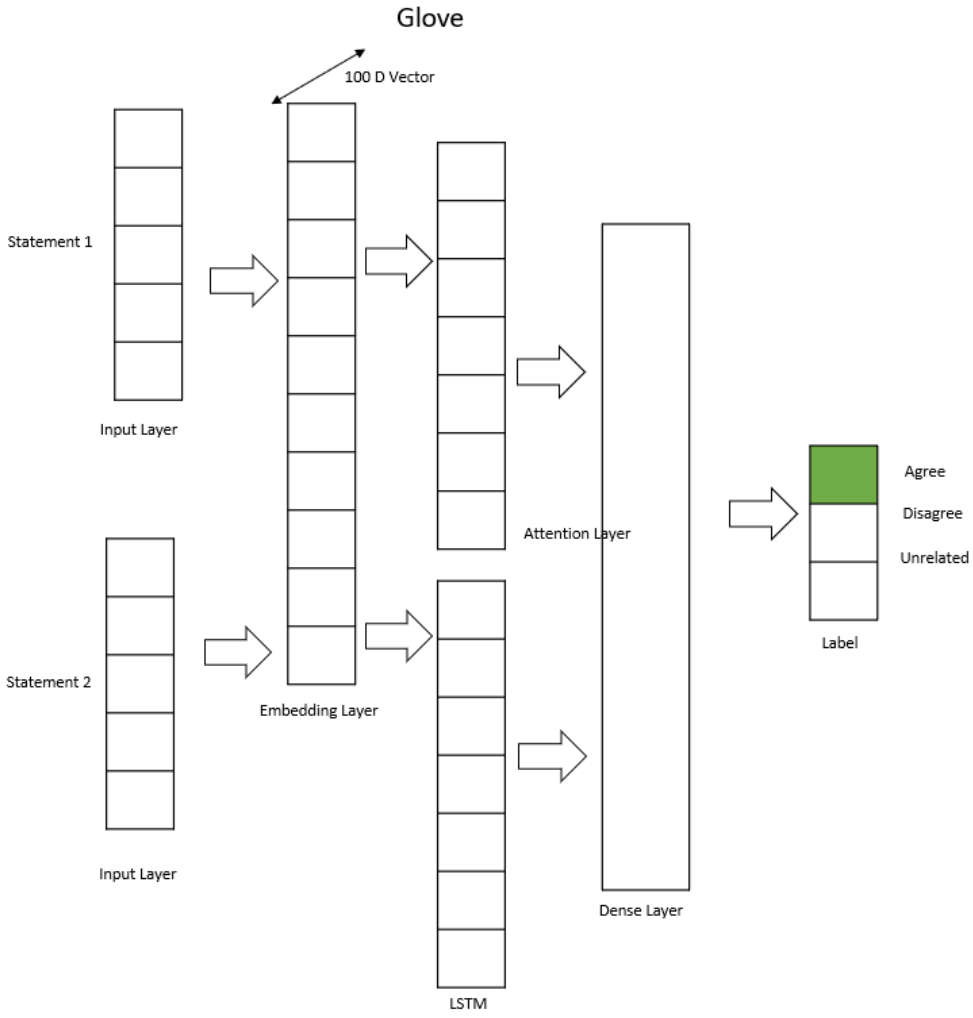

Figure 4.7: Two way representation model

Chapter 5

Discussion

After a thorough hyperparameter tuning on our model, we evaluated the model. our intention of the model is to be effective in efficiently classifying the corpus of statement pairs in a set of experimentsusing the ByteDance dataset.

Hyperparameters	Experiment range	Choice
Activation function	{ReLU, Tanh, Softmax}	Softmax, ReLU
Dropout rate	0-1	0.1
Epochs	50 – 200	50
Optimiser	Adam	Adam
Loss Function	categorical_crossentropy	categorical_crossentropy
L2 penalty	{0.1,0.01,0.001,0.0001}	0.0001

Figure 5.1: Two way representation model

Given the size and our dataset unbalanced nature , overfitting is very easy for a neural network model,i.e on the unseen test set it would be unable to predict accurately. We used earlystoping, dropout and L2 as regularization techniques to overcome overfitting and improve generalization.

Figure5.2 shows the accuracy for the hyperparameter tuning.

Comparison for all models

Machine learning model usually is evaluated based on its accuracy. precision and f1 score also gives a good overview . The classification report has the summary of it precision, recall and accuracy on a test dataset. Figure below is classification report comparison plot of different algorithms . it is evident that attention model outperforms among all.

Accuracy:

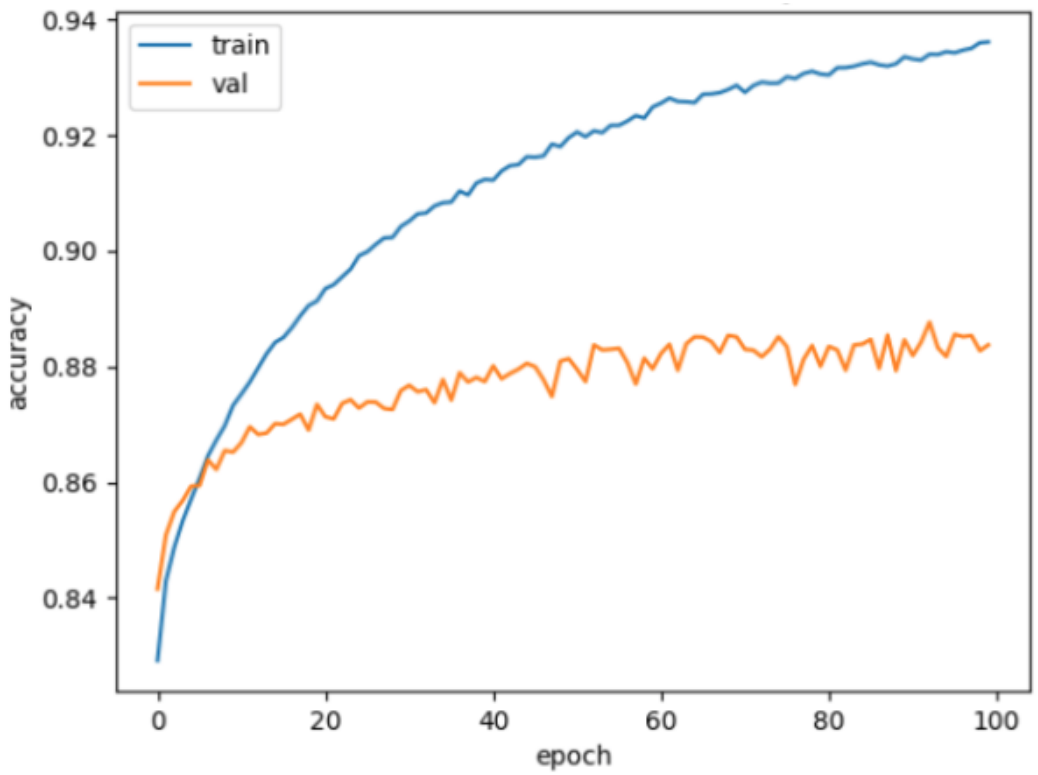

Figure 5.2: Proposed Solution accuracy

Classification Accuracy is what we usually used in term accuracy. It is the ratio of number of correct predictions to the total number of input samples.figure 5.4 **On the Class Imbalance Problem**

We conducted an analysis for balancing the class distribution during training of our model,we aimed to reduce the bias towards the unrelated majority class.for training example The strategy we investigated is as follows: For each training example, if the example is labelled unrelated we included 0.25 prababilty orelse include 1.0 probability. This strategy implies that, the class distribution is balanced slightly between epochs.The confusion matrix for the proposed solution is as below.

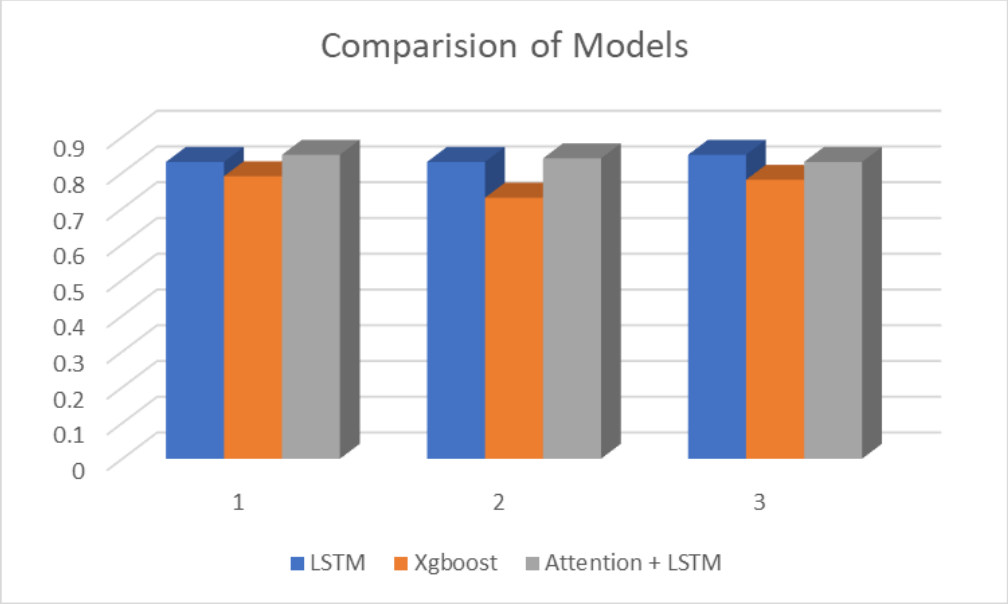

Figure 5.3: Proposed Solution accuracy

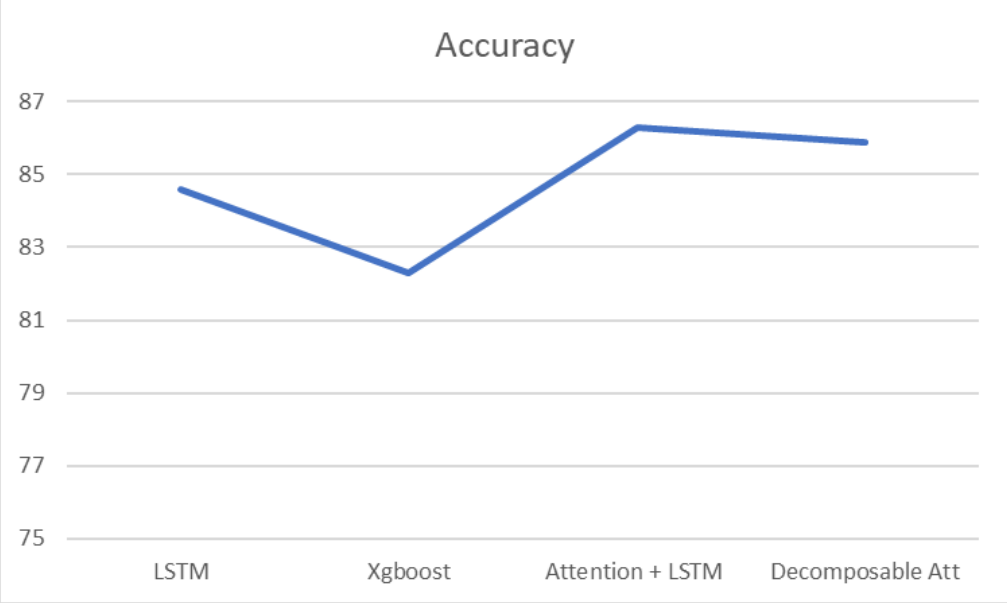

Figure 5.4: accuracy

	Agreed	Disagreed	Unrelated
Agreed	17982	0	12514
Disagreed	47	575	2099
Unrelated	3279	100	69187

Figure 5.5: confusion matrix

Chapter 6

Conclusion and Future Directions

Our experiments showed clearly that deep neural network models with attention and LSTMs significantly outperform on the sequence-based task of fake news detection. Furthermore, we learned that attention seems to have a more modest impact on the context of this specific task. The two way representation allows the network to base its predictions not only on the content but also on their relationship leading to accuracy

Although we are pleased with the classification accuracy delivered by our model on this task, we are interested in considering opportunities for further performance gains. We consider the following for future exploration:

additional training data Given the relatively small size of our training set, especially for the less distribution classes, like agreed and disagreed would help our model learn to generalize better. Of course, it is a hard job because training data provided here must be hand-labeled.

pre-trained embeddings We used 100-dimensional pre-trained GloVe word embeddings, it would be good experimenting to determine whether higher-dimensional vectors might add more accuracy to our models encoding classification capabilities.

additional syntactic features We would be interested in evaluating whether additional language features, such as postagging or named entity labels, would improve the models understanding enough to help it distinguish between classes in the minority of cases where it commits errors today.

The fake news detection problem can be related to Sarcasm detection problem.Interested to see the effect of implementing the existing methods that are more effective in sarcasm detection domain in Fake News detection domain.

List of Figures

List of Tables

Bibliography

[1] Sebastian MÃęland Pedersen. Fake news challenge, 2017. URL http://www.fakenewschallenge.org/.

[2] Dipanjan Das Ankur P. Parikh, Oscar Tackstr. A decomposable attention model for natural language inference/subscribe. *Proc. VLDB Endow.*, 1(1):451–462, 2016. ISSN 2150-8097. doi: https://arxiv.org/pdf/1606.01933.pdf.

[3] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory./subscribe. *Neural Computation archive*, 9(1):1735–1780, 1997. ISSN 2150-8097. doi: https://dl.acm.org/citation.cfm?id=1246450.

[4] Christopher D. Manning Jeffrey Pennington, Richard Socher. Glove: Global vectors for word representation/subscribe. *Neural Computation archive*, 9(1), 2014. doi: https://nlp.stanford.edu/pubs/glove.pdf.

[5] WIKI. FAke news. https://en.wikipedia.org/wiki/Fake$_n ews$, .

[6] Suhang Wang Jiliang Tang Kai Shu, Amy Sliva and Huan Liu. Fake news detectionon social media: A data mining perspective/subscribe. *Neural Computation archive*, 9(1), 2017.

[7] Shlok Gilda. Evaluating machine learning algorithms for fake news detection/subscribe. *Neural Computation archive*, 9(1), 2017.

[8] Yimin Chen Victoria L. Rubin, Niall J. Conroy and Sarah Cornwell. Fake news or truth?using satirical cues to detect potentially misleading news. *Proc. VLDB Endow.*, 1 (1):451–462, 2016. ISSN 2150-8097. doi: https://www.aclweb.org/anthology/W16-0802.

[9] William Ferreira Andreas Vlachos. Emergent: a novel data-set for stance classification. *Proc. VLDB Endow.*, 1(1):451–462, 2016. ISSN 2150-8097. doi: https://www.aclweb.org/anthology/W16-0802.

[10] HONG-NING DAI 1 HAO WANG 2 JUNHAO ZHOU1, YUE LU1. Sentiment analysis of chinese microblog based on stacked bidirectional lstm. *Proc. VLDB Endow.*, 1(1):

451–462, 2016. ISSN 2150-8097. doi: https://www.henrylab.net/wp-content/uploads/2019/04/08667413.pdf.

[11] J.; Reinartz-K.; Bevendorff J.; Potthast, M.; Kiesel and Stein. A stylometric inquiry into hyperpartisan and fake news. *Neural Computation archive*, 15(1), 2017.

[12] Asif Ekbal Pushpak Bhattach Detection: A.Arjun Roy, Kingshuk Basak. A deep ensemble framework for fake news detection a/subscribe. *Neural Computation archive*, 9(1), 2018.

[13] Bernard J Jansen Kam-Fai Wong Mitra, Sejeong Kwon. Detectingrumors from microblogs with recurrent neural network/subscribe. *Neural Computation archive*, 9(1), 2016.

[14] Sahil Chopra and Saachi jain. Towards automatic identification of fake news:headline-article stance detection with lstm atttention/subscribe. *Neural Computation archive*, 9 (1), 2017.

[15] Giorgos Myrianthous-Jiashu Pu Xiaoxuan Wang sifiers James Thorne, Mingjie Chen and Andreas. Fake news stance detection using stackedensemble of classifier. In *In Proceedings of the 2017 EMNLP Workshop: Natural LanguageProcessing meets Jou*, pages 519–526, New York, NY, USA, 2017. Aclweb. ISBN 978-1-59593-597-7. doi: https://aclweb.org/anthology/papers/W/W17/W17-4214/.

[16] Jin Yea Jang-Svitlana Volk Hannah Rashkin, Eunsol Choi. Truthof varying shades: Analyzing language in fake news and politics. In *itical fact-checking. In Pro-ceedings of the 2017 Conference on Empirical Methods in Natural La*, pages 519–526, New York, NY, USA, 2017. Aclweb. ISBN 978-1-59593-597-7. doi: https://aclweb.org/anthology/D17-1317.

[17] Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural machine translation by jointly learning to align and translate/subscribe. *Neural Computation archive*, 9(1), 2014.

[18] Victor Zhong Caiming Xiong. Dynamic coattention networks for question answering/-subscribe. *Neural Computation archive*, 9(1), 2016.

[19] Karl Moritz Hermann-TomÃąÅą KoÄĎiskÃ¡ Phil Blunsom Tim RocktÃďschel, Edward Grefenstette. Reasoning about entailment with neural attention. *Proc. VLDB Endow.*, 1(1):451–462, 2015. ISSN 2150-8097. doi: https://arxiv.org/abs/1509.06664.

[20] Bing Xiang Bowen Zhou Wenpeng Yin, Hinrich SchÃijtze. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Proc. VLDB Endow.*, 1(1): 451–462, 2015. ISSN 2150-8097. doi: https://arxiv.org/abs/1512.05193.

[21] Oscar Inference Ankur P. Parikh. A decomposable attentionmodel for natural language/-subscribe. *Neural Computation archive*, 9(1), 2016.

[22] Atul. machine learning categories, 2014. URL https://www.edureka.co/blog/what-is-machine-learning/.

[23] Jason. machine learning categories, 2014. URL https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

[24] Atul. neural network, 2014. URL https://blog.exxactcorp.com/lets-learn-the-difference-between-a-deep-learning-cnn-and-rnn/.

[25] WIKI. bag of words. https://en.wikipedia.org/wiki/Bag-of-words$_model$/, .

[26] Atul. neural network, 2014. URL http://www.tfidf.com/.

[27] Christopher D. Manning Jeffrey Pennington, Richard Socher. Fake news challenge, 2014. URL https://nlp.stanford.edu/projects/glove/.

[28] Atul. neural network, 2014. URL https://blog.algorithmia.com/introduction-to-loss-functions/.

[29] Y. Bengio Goodfellow and A. Courville. *Deep learning*. MIT Press, Cambridge, MA, USA, second edition, 2016. ISBN 0-262-03293-7.

[30] Nitish Srivastava. A simple way to prevent neural networks from overfitting /subscribe. *Neural Computation archive*, 15(1), 2014.

[31] Christopher D. Manning Jeffrey Pennington, Richard Socher. class imbalance, 2017. URL https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a.

[32] Christopher D. Manning Jeffrey Pennington, Richard Socher. accuracy/precision, 2017. URL https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

[33] E. Barrenechea H. Bustince M. Galar, A. FernÃąndez. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approache/subscribe. *Neural Computation archive*, 15(1), 2011.

[34] Christopher D. Manning Jeffrey Pennington, Richard Socher. accuracy/-precision, 2017. URL https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

[35] Dipanjan Das Ankur P. Parikh, Oscar Tackstr. A decomposable attention model for natural language inference/subscribe. *Proc. VLDB Endow.*, 1(1):451–462, 2016. ISSN 2150-8097. doi: https://arxiv.org/pdf/1606.01933.pdf.

[36] shagunsodhani. attention, 2017. URL https://shagunsodhani.com/papers-I-read/A-Decomposable-Attention-Model-for-Natural-Language-Inference.

[37] Siu Cheung HuiÏĘ Jian SuÎt Yi TayâĂă, Luu Anh TuanÏĹ. Reasoning with sarcasm by reading in-between /subscribe. *Neural Computation archive*, 15(1), 2018.