TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Khwopa College Of Engineering
Libali, Bhaktapur
Department of Computer Engineering

A PROPOSAL ON
FAKE NEWS DETECTION USING NLP and LSTM

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF COMPUTER ENGINEERING

Submitted by

Ronish Shrestha	KCE075BCT032
Roshan Shrestha	KCE075BCT033
Rubin Baidhya	KCE075BCT034
Sairush Tamang	KCE075BCT036

Under the Supervision of
Department Of Computer Engineering

Khwopa College Of Engineering
Libali, Bhaktapur
2021-22

Certificate of Approval

The undersigned certify that the mini project entitled **"Fake News Detection using NLP and LSTM"** submitted by Ronish Shrestha, Roshan Shrestha, Rubin Baidhya and Sairush Tamang to the Department of Computer Engineering in partial fulfillment of requirement for the degree of Bachelor of Engineering in Computer Engineering. The project was carried out under special supervision and within the time frame prescribed by the syllabus.

....
Er.
(Project Supervisor)

....
Er. Dinesh man Gothe
Head of Department
Department of Computer Engineering, KhCE

Copyright

Acknowledgement

We would like to express our deepest gratitude to our principal **Er. Sunil Duwal** for providing us with this opportunity to do this project.

Similarly, we would like to acknowledge our teachers **Er. Dinesh Man Gothe**, **Er. Bindu Bhandari** and **Er. Sajit Pyakurel** for guiding and encouraging us to initiate our project. We are deeply indebted to them for sharing us with their experiences that have helped us in many ways throughout our Bachelor's program.

Finally, we would like to thank each and everyone who have directly and indirectly assisted us in starting our project.

Ronish Shrestha	KCE075BCT032
Roshan Shrestha	KCE075BCT033
Rubin Baidhya	KCE075BCT034
Sairush Tamang	KCE075BCT036

Abstract

Starting the day with the news keeps people updated on current issue of the world. Social media has been the fastest, easiest and rapid way to circulate news.

However the most preferred news source comes with the great risk of exposure to “fake news” that is written intentionally to mislead the readers. Fake news not only mislead the readers but also creates a devastating impact so it is a major concern. Thus the increasing fake news globally tends to encroach human right and public safety.

A person needs to have vast knowledge about the current issue to discover whether a news is real or fake. So we propose a system to detect the fake news using Natural Language Processing (NLP) and long short term memory (LSTM).

Based on the headlines and the content of the news the proposed system recognizes whether the news is fake or real in short amount of time.

Keywords: *Fake news, Natural language processing, Long short term memory, Neural networks*

Contents

List of Figures

List of Symbols and Abbreviation

ML	Machine Learning
NLP	Natural Language Processing
LSTM	Long Short Term Memory
AI	Artificial Intelligence
CNN	Convolutional Neural Network
RNN	Recurrent Nerual Network
TF-IDF	Term Frequency Inverse Document Frequency
GRU	Gated Recurrent Unit
PCA	Principal Component Analysis
IDS	Intrusion Detection System
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
SDLC	Software Development Life Cycle

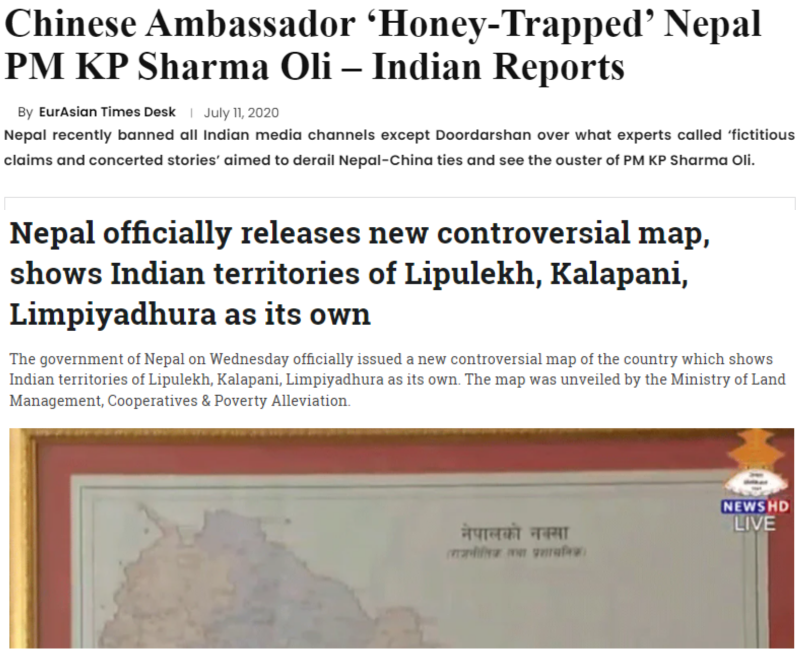
Chapter 1

Introduction

1.1 Background

With tremendous amount of data flowing over internet, most of them are found fake but still people prefer the internet source as it is the easiest means. Unfortunately the open source has become the platform to release news without effective supervision. Fake news are delivered cleverly in a traditional method that implies to be truthful but contain misleading information. By utilizing different philosophies, realizing the correctness of the news is a fascinating issue [7]. As defined by the New York Times, fake news is "made up stories written to deceive" [8].

These types of news are based on personal opinions and are generated to attract the audience, to influence their understanding and decisions, to multiply the income by clicking and to affect any major events [10]. For example, During the COVID-19 pandemic, different conspiracy theories about its origin (such as 5g cellular network) and the preventive measures like sipping water every 15 minutes, taking cocaine. On the top of that, false news is used as props increasing harmful consequences in the field of business, stock market and politics. Back in July 2020 some Indian news media fabricated news report by alleging a relationship between former Prime Minister K P Sharma Oli and Chinese Ambassador Hou Yanqi. Fig 1 shows few examples of fake news spread over social media.



Chinese Ambassador 'Honey-Trapped' Nepal PM KP Sharma Oli – Indian Reports

By EurAsian Times Desk | July 11, 2020

Nepal recently banned all Indian media channels except Doordarshan over what experts called 'fictitious claims and concerted stories' aimed to derail Nepal-China ties and see the ouster of PM KP Sharma Oli.

Nepal officially releases new controversial map, shows Indian territories of Lipulekh, Kalapani, Limpiyadhura as its own

The government of Nepal on Wednesday officially issued a new controversial map of the country which shows Indian territories of Lipulekh, Kalapani, Limpiyadhura as its own. The map was unveiled by the Ministry of Land Management, Cooperatives & Poverty Alleviation.

Figure 1.1: Examples of few fake news

Such news spread negative impact and hampers the public emotionally. Without the proper understanding and the written facts about the news it is tedious to claim a news as fake and also the conventional method to check the claim is time consuming and expensive. Hence it is most to develop methods for detecting fake news. In this proposed model news are classified from the perspective of NLP automatically based on the agreement between the headline and the body

1.2 Problem

The daily life of people is hugely influenced by information. Most of the people plan their daily routine based on the information they obtain from various sources. This shows our reliance on news and also presents the importance of real news on daily basis.

Although a simple piece of news carry huge importance in every social, economic, political and environmental fields, the credibility of that piece of news can cause great turmoil in the world. A true news will not create any negative impacts on our society but if fake news is spread across the society and country then it will very badly impact the society, country and the people.

People consume news from various medias and there are presently all sorts of news available in online sources. Some people use such online platforms to spread various fake news for their own gains. This creates confusion among the people and create havoc among the people. People themselves may not be able to classify whether the news is real or fake and blindly follow such news. This is a great problem that we are facing nowadays. But now we can classify a news as real or fake using artificial intelligence and deep learning. However most of the current existing fake news detection system only classify news based on the headline and the content. Since news can also be an image or video, these format are not processed in existing systems.

Integrating AI can help us identify fake news and filter them before they create confusion among the people.

1.3 Objectives

The major objective of the project ”Fake News Detection” is to reduce the effort and time to detect fake news and help stop spreading it. With emerging data in internet it is necessary to classify a news as fake or real as it is a major problem. In the field of AI, Fake news detection has been one of the emerging topic. Despite the researchers are still thriving to improve the accuracy. Our study, on the other hand, uses Natural Language Processing and Neural Networking to boost accuracy while reducing human effort.

1.4 Scope and Limitations

Fake news creates confusion and disbelief among the people. They may appear in different forms such as click-baits, misleading headlines, fabricated content and biased news.

So implementing fake news detecting techniques might help in minimizing the spreading of such disinformation and also protect people’s right to true information.

But the advancing technology is posing more and more challenges in the detection of fake news, as the influence of social media on people grows.. Similarly the detection of false news is becoming more difficult, as news is becoming more elegantly titled as if it were true.

Similarly, the presently available machine learning and deep learning techniques still face many challenges to detect the fake news as the content are being planned in such a way that they resemble truth.

Also, the irregular and unpredictable data may lead the prediction model to make mistakes during detection.

Chapter 2

Literature Review

Fake reviews and fake news both are responsible for spreading misconceptions and disbeliefs among the people, according to the article ”Detecting opinion spams and fake news using text classification” [2]. The article has also emphasized that detecting fake news is even harder than the detection of fake reviews. It has categorized fake news into 3 groups. The first being false news, second as fake satire news and lastly poorly written news articles. Here, the authors have introduced models for detecting both fake news and fake reviews. The authors have also mentioned them being the first to detect both the fake news and fake reviews at the same time. The authors have mentioned that they had used TF-IDF as the feature extraction model for their datasets. The prediction model was designed using six different classifiers by checking their respective accuracy and efficiency.

[1] has proposed a method to measure the user's credibility in social media. The authors have proposed CredRank algorithm for measuring user credibility in social media. This algorithm is supposed to detect user's online behavior and measure the credibility of user accordingly. Here, Abbasi, Mohammad-Ali and Liu, Huan have taken into consideration the Arab Spring movements in social media sites, and shown how the users can spread misinformation in the society.

The research article [5] has shown the use of Naive Bayes classifier for fake news detection. The authors used the data from facebook news posts for implementing the Naive Bayes classifier and also confirmed that they got decent accuracy of around 74% using the classifier. They even stressed on the point that using a complex model can result in even greater accuracy and efficiency.

Gilda [4] has explored the usage of neural networking algorithms for fake news detection in his paper. The author experimented his data in various classifiers and calculated their respective accuracies. The author has also stressed that Stochastic Gradient Descent model identifies non-credible sources with higher accuracy with respect to other models.

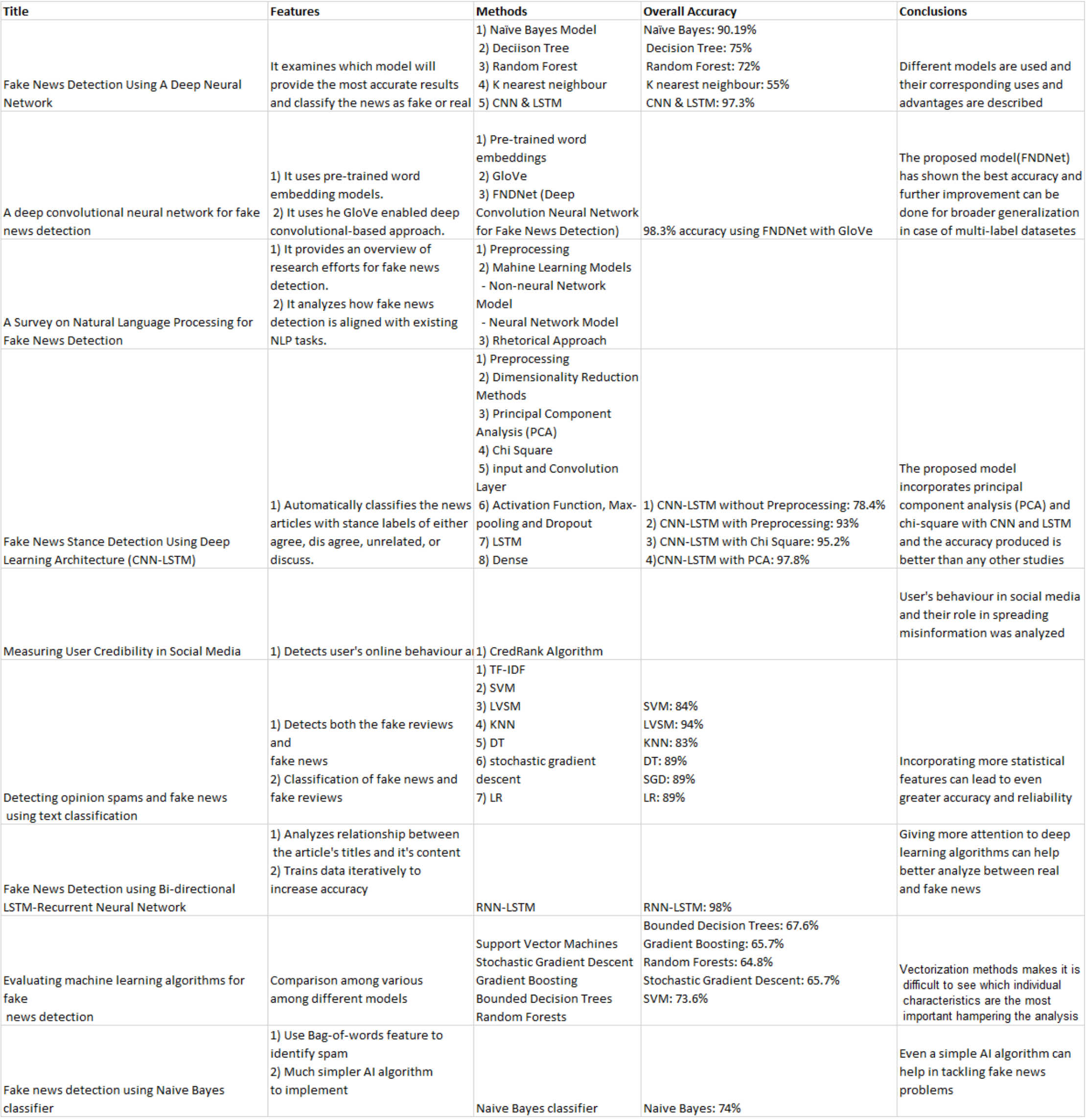
Bahad et. al. [3] has presented bi-directional LSTM model for fake news detection as a superior model as compared to other models namely uni-directional LSTM, CNN and vanilla RNN. The authors have also expressed that the traditional text mining and machine learning techniques are inadequate in handling the large and complex datasets available in present days. The proposed model analyzes the relationship between the article's titles and it's content to detect whether the article is real or fake. The proposed RNN-LSTM trains the data iteratively for improving the accuracy. It was concluded by the authors that the LSTM-RNN classifier is suitable long-range semantic dependency based classification. Also, the model was very capable in working with both balanced and imbalanced high dimensional news data set.

In article [7] various Machine Learning models like Naive Bayes, K nearest neighbour, Decision Tree, Random Forest and Deep Learning networks(CNN, LSTM, GRU) are used. He has also explored and used features like n-gram, TF-IDF. All these models were used and their corresponding accuracies were pointed out due to which evaluation becomes easier. On comparing all these models, CNN LSTM using together was found to have the highest accuracy of 97.3%.

Rohit kumar kaliyar et. al [8] has proposed FNDNet, a deep convolutional neural network for fake news detection. To analyze FNDNet, GloVe is used as pre-trained word embedding which is unidirectional in training. They have also used various machine learning as well as deep learning algorithms for classification. With the accuracy of 98.3% the authors have stated that their proposed model provides ”state-of-art” result for predicting fake news. Despite the high performance of their classifiers, they have realized that there is still room for improvement. The authors have concluded that their further plan is to use hybrid approach for fake news classification as it provides broader generalization which is lacking in their model. FNDNet with Glove has proved to be the highest accuracy yielding algorithm amongst the other papers so far. So it is convincing that it needs to be implemented in the field of fake news detection and classification.

Oshikawa, Ray. Jing, Qian. Wang, William. Y. in their survey [10] has revealed the importance of automatic fake news detection. This survey for sure has emphasized on the related problems regarding fake news detection and has proposed a method using NLP. To compare results they have mainly focused on three datasets: LIAR, FEVER, and FAKENEWSNET. The authors have suggested to examine if hand-crafted features may be combined with neural network, as well as the suitable use of non-textual data and extending the way of verification with contents.

[11] The authors have proposed a fake news stance detection model that classifies the news articles with stance labels of either agree, disagree, unrelated, or discuss. The proposed model uses Principal Component Analysis (PCA) and Chi-square with CNN-LSTM. PCA and Chi-square together perform component level analysis and obtain the reduced feature set which are passed to CNN-LSTM model. This model produces a promising accuracy by scoring 97.8%. However it is seen that while preserving the high performance of classifiers by using Dimensionality reduction techniques(for removing redundant features, noisy and irrelevant data) it can reduce the number of features. So in order to boost the performance of their model, the authors have mentioned that their future work will analyze different textual features, tree-based learning for simple approach and validate the performance of their proposed model for larger datasets. Amongst all other papers reviewed, this model gives the second highest accuracy for fake news detection.



Title	Features	Methods	Overall Accuracy	Conclusions
Fake News Detection Using A Deep Neural Network	It examines which model will provide the most accurate results and classify the news as fake or real	1) Naïve Bayes Model 2) Deciison Tree 3) Random Forest 4) K nearest neighbour 5) CNN & LSTM	Naïve Bayes: 90.19% Decision Tree: 75% Random Forest: 72% K nearest neighbour: 55% CNN & LSTM: 97.3%	Different models are used and their corresponding uses and advantages are described
A deep convolutional neural network for fake news detection	1) It uses pre-trained word embedding models. 2) It uses he GloVe enabled deep convolutional-based approach.	1) Pre-trained word embeddings 2) GloVe 3) FNDNet (Deep Convolution Neural Network for Fake News Detection)	98.3% accuracy using FNDNet with GloVe	The proposed model(FNDNet) has shown the best accuracy and further improvement can be done for broader generalization in case of multi-label datasetes
A Survey on Natural Language Processing for Fake News Detection	1) It provides an overview of research efforts for fake news detection. 2) It analyzes how fake news detection is aligned with existing NLP tasks.	1) Preprocessing 2) Mahine Learning Models - Non-neural Network Model - Neural Network Model 3) Rhetorical Approach		
Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)	1) Automatically classifies the news articles with stance labels of either agree, dis agree, unrelated, or discuss.	1) Preprocessing 2) Dimensionality Reduction Methods 3) Principal Component Analysis (PCA) 4) Chi Square 5) input and Convolution Layer 6) Activation Function, Max-pooling and Dropout 7) LSTM 8) Dense	1) CNN-LSTM without Preprocessing: 78.4% 2) CNN-LSTM with Preprocessing: 93% 3) CNN-LSTM with Chi Square: 95.2% 4)CNN-LSTM with PCA: 97.8%	The proposed model incorporates principal component analysis (PCA) and chi-square with CNN and LSTM and the accuracy produced is better than any other studies
Measuring User Credibility in Social Media	1) Detects user's online behaviour a	1) CredRank Algorithm		User's behaviour in social media and their role in spreading misinformation was analyzed
Detecting opinion spams and fake news using text classification	1) Detects both the fake reviews and fake news 2) Classification of fake news and fake reviews	1) TF-IDF 2) SVM 3) LVSM 4) KNN 5) DT 6) stochastic gradient descent 7) LR	SVM: 84% LVSM: 94% KNN: 83% DT: 89% SGD: 89% LR: 89%	Incorporating more statistical features can lead to even greater accuracy and reliability
Fake News Detection using Bi-directional LSTM-Recurrent Neural Network	1) Analyzes relationship between the article's titles and it's content 2) Trains data iteratively to increase accuracy	RNN-LSTM	RNN-LSTM: 98%	Giving more attention to deep learning algorithms can help better analyze between real and fake news
Evaluating machine learning algorithms for fake news detection	Comparison among various among different models	Support Vector Machines Stochastic Gradient Descent Gradient Boosting Bounded Decision Trees Random Forests	Bounded Decision Trees: 67.6% Gradient Boosting: 65.7% Random Forests: 64.8% Stochastic Gradient Descent: 65.7% SVM: 73.6%	Vectorization methods makes it is difficult to see which individual characteristics are the most important hampering the analysis
Fake news detection using Naive Bayes classifier	1) Use Bag-of-words feature to identify spam 2) Much simpler AI algorithm to implement	Naive Bayes classifier	Naive Bayes: 74%	Even a simple AI algorithm can help in tackling fake news problems

Figure 2.1: Synthesis Matrix

Chapter 3

Requirement Analysis

3.1 Software Requirements

3.1.1 Python

Python is an interpreted high-level general-purpose programming language. Its design emphasizes code readability by using a lot of indentation. The language features and object-oriented approach of python are designed to help programmers write clear, logical code for both small and large-scale projects.

Keras

Keras is a python-based deep learning framework that is open source. Francois Chollet, a Google artificial intelligence researcher, came up with the idea. Keras adheres to best practices for lowering cognitive load, such as providing uniform and simple APIs, limiting the number of user activities required for typical use cases, and providing clear and actionable error messages. It comes with a lot of documentation and developer instructions. Keras is presently used by Google, Square, Netflix, Huawei, and Uber, among others.

3.1.2 Machine Learning, Deep Learning and Neural Networks

The science of making computers to act without being explicitly programmed is known as machine learning. Self-driving cars, realistic speech recognition, successful web search, and a much developed understanding of the human genome have all been made possible by machine learning. the last decade.

Deep Learning is a machine learning and artificial intelligence(AI) technique that mimics how humans acquire knowledge. Data science, which covers statistics and predictive modeling, incorporates deep learning as a key component. Deep learning is highly useful for data scientists who are responsible with gathering, analyzing, and interpreting massive amounts of data; it speeds up and simplifies the process.

A neural network is a set of algorithms that attempts to recognize underlying relationships in a batch of data using a method that mimics how the human brain works. Neural networks, in this context, refer to systems of neurons that can be organic or artificial in nature. Because neural networks can adapt to changing input, they can produce the best possible outcome without requiring the output criteria to be redesigned.

Natural Language Processing(NLP)

Natural language processing (NLP) is a sub-field of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Long Short-Term Memory(LSTM)

Long short-term memory(LSTM) is a deep learning architecture that uses an artificial recurrent neural network (RNN). There are feedback connections in the LSTM. it can handle not only individual data points(such as photos), but also unsegmented, connected handwriting identification, speech recognition, and anomaly detection in network traffic or IDS (intrusion detection systems) can all benefit from LSTM. A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. the three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time intervals.

3.2 Functional Requirements

The main goal of this project is to determine whether or not the contents of a news story presented are genuine. So, followings can be listed as the functional requirements of our Fake News Detection Project:

- The system should be adaptable and elaborated from the point of view of the users.
- The system should able to distinguish the types of news content.
- The system should figure out the news that are completely fake or not genuine.

3.3 Non-Functional Requirements

The following non-functional requirements should be met by our project on fake news identification.

3.3.1 Reliability

After completion, the project will be extremely dependable. The system will be able to run without fail in a specific environment.

3.3.2 Maintainability

The maintenance of the finished system will be easily feasible. Also we can easily maintain our system for future enhancements. By adding additional examples to the database with which we train our system, the system's ability to identify can be easily upgraded.

3.3.3 Performance

Performance of the system depends on how fast the system gives the result. After its completion, the system will be able to detect fake news in a very less time.

3.3.4 Accuracy

The technology should be able to tell whether the news being read is phony or trustworthy. The system will be capable of accurately detecting fake news.

3.4 System Requirement

- Working mobile or PC with internet connection

Chapter 4

Feasibility Study

A feasibility study is an examination of a project's relevant variables, such as economic, technical, legal, and scheduling, with the goal of successfully completing the project. It considers the advantages and disadvantages of a project before committing time and money to it. It determines whether or not the proposed system is incompatible with the requirements. For instance, operation-wise, spacing, time, software legality, and so on. A project feasibility study is a report that evaluates the frames of analysis for a certain project in great depth.

4.1 Time Feasibility

This project's development cycle is estimated to take 5 months or perhaps longer. We want to complete the assignment as soon as possible in order to make it more practical while still meeting the deadline.

4.2 Space Feasibility

We're attempting to keep this project as minimal as possible in order to save disk space. To reduce space complexity, it will be preferable to use fewer libraries and frameworks.

4.3 Technical Feasibility

All of the data sets that are required for our project are easily available in the internet. We can complete the job using currently available tools. Open sourced and freely available tools will be adequate to meet the system's technical feasibility requirements.

4.4 Economic Feasibility

We want to employ the majority of open-source tools, libraries, frameworks, and data sets available on the internet. As a result, there is no expense. There is no investment of any budget in this project

4.5 Operational Feasibility

We are proposing to develop a web-based fake news detection, mainly focused on more user-friendly and informative environment where user will able to use the system in more convenient fashion.

Chapter 5

Development Resources

5.1 Datasets

The followings are the datasets that will be used in our Fake News Detection Project:

- Kaggle [6] : It is the website where we extracted the data of fake and true news.
- Buzz Feed News [9] : This little dataset was created with the help of Facebook. The required dataset is fact-checked by five Buzzfeed journalists. Only the headlines and text of 2282 posts are included.

5.2 Tools and Materials

- Python
- Git
- Jupyter Notebook
- HTML and CSS

Chapter 6

System (or Project) Design and Architecture

6.1 Structure of System

The basic structure of our application is very simple. It is a client-server based architecture as shown in figure 6.1.

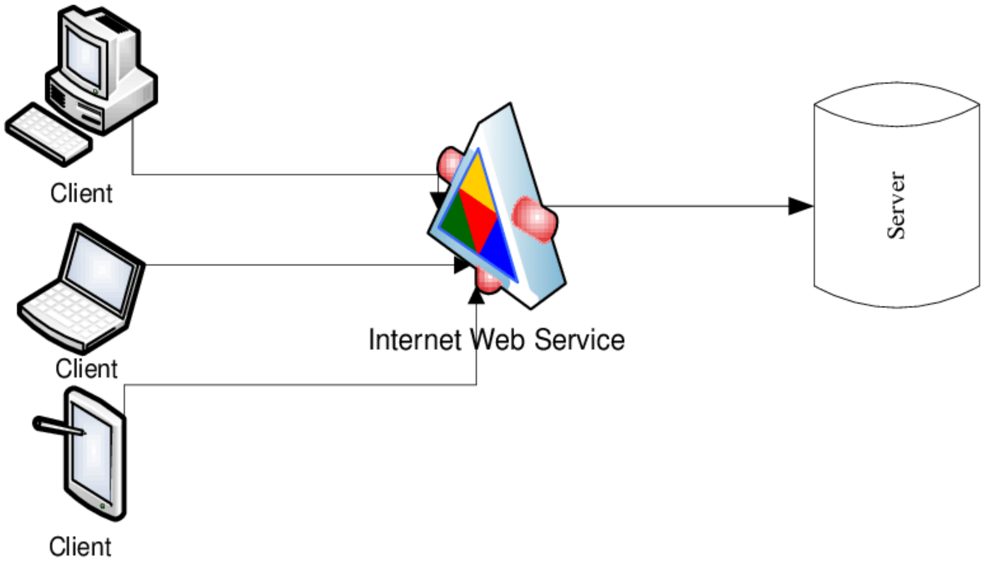


Figure 6.1: Client Server Architecture

The user delivers the piece of news or the url of any portal website's news article to the server from the frontend application, and the server simply returns if the news article is fake or true. We can detect false news in the backend or in the frontend. The model developed using python and its libraries will be kept on the server or in any cloud-based storage facilities. All news articles will be stored in the database, along with a property that indicates if the article is fake or not.

6.2 System Block

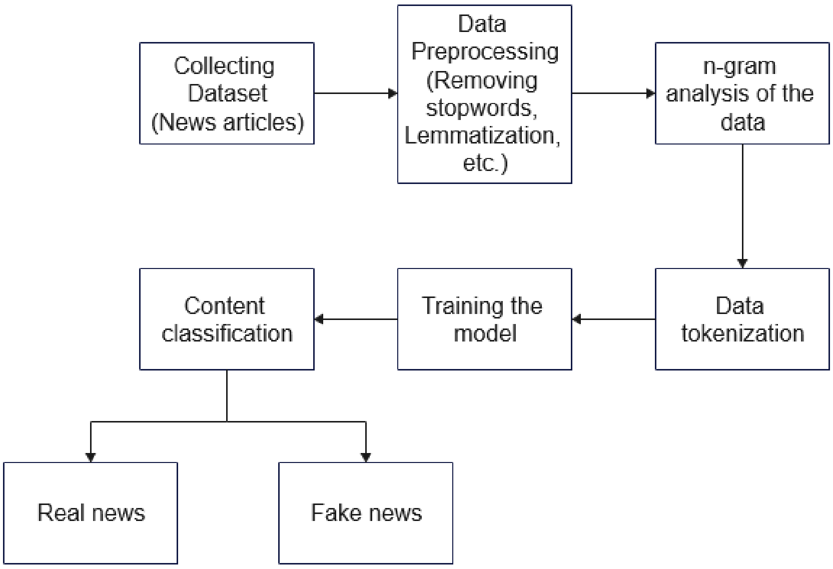

Collecting Dataset (News articles)
Data Preprocessing (Removing stopwords, Lemmatization, etc.)
n-gram analysis of the data
Data tokenization
Training the model
Content classification
Real news
Fake news

Figure 6.2: System Block Diagram

6.3 USE CASE DIAGRAM

The system will be trained using the dataset and the admin will be able to add new materials to the dataset to keep the system updated and more reliable for the future use as well. The user can make use of the system's facility to determine whether the news he is reading is fake or trustworthy.

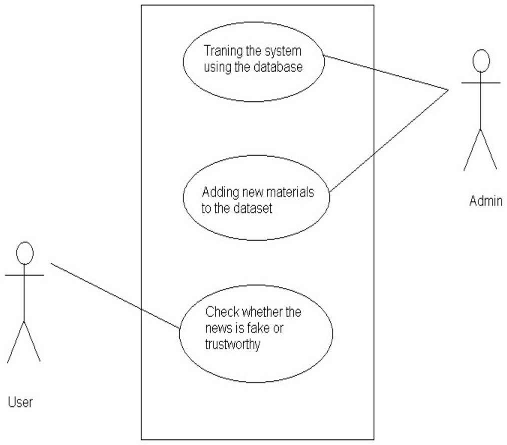

Traning the system using the database
Adding new materials to the dataset
Check whether the news is fake or trustworthy
Admin
User

Figure 6.3: Use Case Diagram

6.4 Sequence Diagram

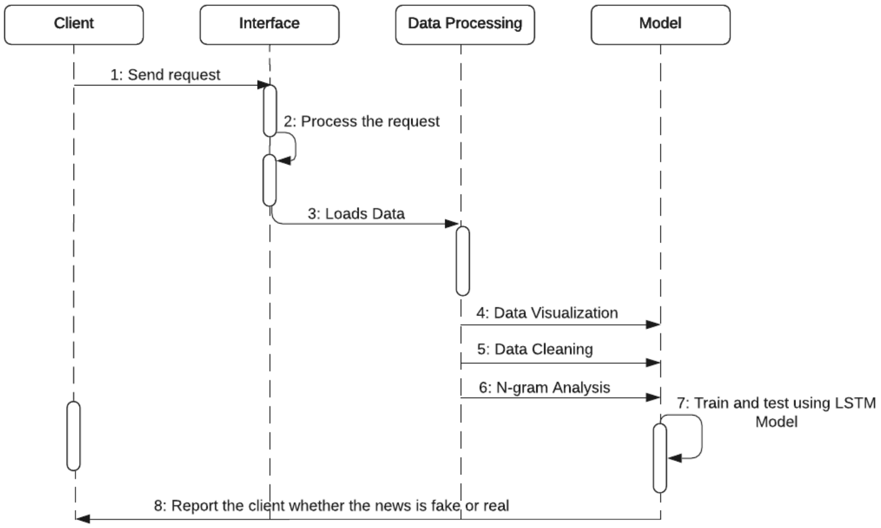


Figure 6.4: Sequence diagram

Chapter 7

Methodology

7.1 Data Collection

The datasets we are going to use are downloaded from [6] and [9],which will be used for modeling our system for the detection.

7.2 Dataset Preprocessing

The datasets that are mentioned in the Chapter 5 consists of the fake and genuine news articles. These datasets might contain duplicate news article. So, we are planning to preprocess the datasets to remove all the duplicate news. We will be using python library i.e beautiful soup 4 for web scraping and removing the html contents from the article. A generic preprocessing function is developed that removes punctuation and non-letter characters from each news article of the datasets.And, the datasets items are converted to lowered case. In addition, an n-gram word-based tokenizer was developed to slice the document text into n-gram chunks.

After tokenizing the data, the tokens must be converted into a standard format. Stemming is the process of returning words to their original form while reducing the number of word categories or classes in the data.In contrast to stemming, lemmatization looks beyond word reduction and considers a language's full vocabulary to apply a morphological analysis to words. The lemma of 'was' is 'be' and the lemma of 'mice' is 'mouse'. Also,Multiple variants of the same word are lemmatized to return to their common root, such as 'coming' and 'comes' into 'come'. Lemmatization is typically seen as much more informative than simple stemming. This will assist us in reducing the size of the actual data by deleting any unnecessary information.

7.3 N-Gram Analysis

In the fields of language modeling and NLP, N-gram modeling is a prominent feature identification and analysis approach. Word-based and character-based n-grams are the most commonly utilized n-gram models in text categorization. We employ a word-based n-gram to capture the document's context and produce features to classify it in this paper. to distinguish between fake and genuine news articles, we create a basic n-gram based classifier. The objective is to produce various sets of n-gram frequency profiles to represent fake and true news stories.

We looked at the effect of n-gram length on the accuracy of several classification algorithms using numerous baseline n-gram features based on words.

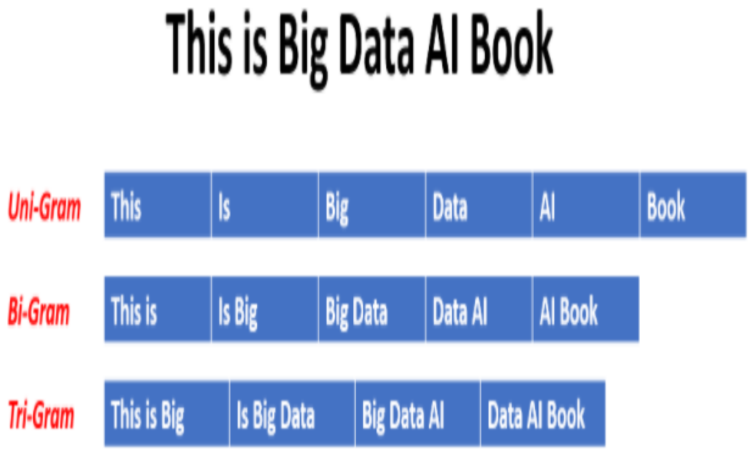
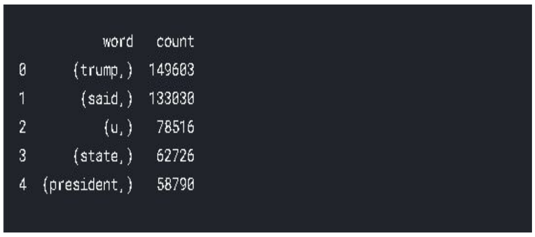


Figure 7.1: N-gram

Uni-gram Analysis

A uni-gram is a one word sequence. It is a set of words selected only one at a time.

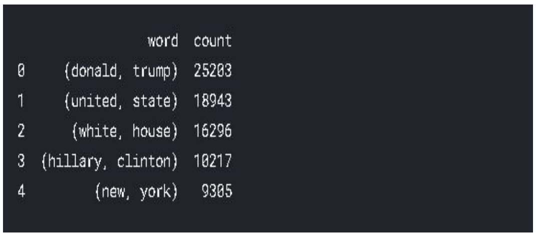


```
           word   count
0      (trump,)  149603
1       (said,)  133030
2          (u,)   78516
3      (state,)   62726
4  (president,)   58790
```

Figure 7.2: Uni-gram Data Set

Bi-gram Analysis

A bi-gram is a two word sequence.

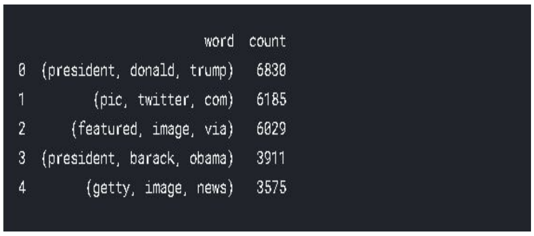


```
                 word  count
0     (donald, trump)  25203
1     (united, state)  18943
2      (white, house)  16296
3  (hillary, clinton)  10217
4         (new, york)   9305
```

Figure 7.3: Bi-gram Data Set

Tri-gram Analysis

A tri-gram is a three words sequence.



```
                         word  count
0  (president, donald, trump)   6830
1        (pic, twitter, com)    6185
2     (featured, image, via)    6029
3  (president, barack, obama)   3911
4       (getty, image, news)    3575
```

Figure 7.4: Tri-gram Data Set

7.4 Modeling

7.4.1 Train-Test Split

After the n-gram analysis, the input features are used to train LSTM model. Each dataset is split into two parts: training and testing, with 70/30 ratio in each.

7.4.2 Tokenizing

Tokenizing can also be known as feature extraction. In tokenization each word are represented by a number. This step is crucial. To extract and build the features for our applications, we will use keras approach. After the original word is converted to number its mapping is stored in word index property of the tokenizer.

7.4.3 Training LSTM Model

We must create a long chain-like sequence structure for our data while retaining earlier input knowledge. Since Traditional neural networks can't recall or keep track of everything that has passed through them, LSTM is the most appropriate solution for this task. LSTM is a building blocks for the layers of a recurrent neural network.

7.5 SOFTWARE DEVELOPMENT APPROACH

We will develop the application in incremental and iterative way. As a result, the SDLC model we intend to employ is **Agile SDLC model**. This is for the creation of a functional prototype before the mid-defense period. As a result, we will have developed a simple web prototype of our application in the first portion of the project.

Gantt Chart

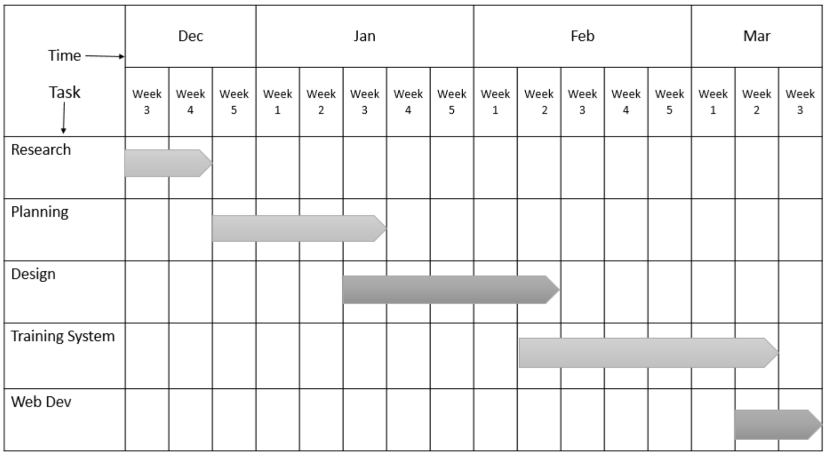


Figure 7.5: Gantt Chart

Chapter 8

Expected Outcomes

We have expected to develop a web application of fake news detection. The final product is expected to predict any news articles as real or fake. We are expecting the output model to be as accurate as possible with an accuracy of over 90%. We expected the performance of our model to be quick as possible. The interface will be simple and attractive so that any user can easily use it.

The user will be provided with an interface where he/she can input any news and the processing will be done in backend. Thus finally, the output will be displayed showing whether the news is real or fake.

Bibliography

[1] Mohammad-Ali Abbasi and Huan Liu. “Measuring User Credibility in Social Media”. In: *Social Computing, Behavioral-Cultural Modeling and Prediction*. Ed. by Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 441–448. ISBN: 978-3-642-37210-0.

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. “Detecting opinion spams and fake news using text classification”. In: *Security and Privacy* 1.1 (2018), e9.

[3] Pritika Bahad, Preeti Saxena, and Raj Kamal. “Fake News Detection using Bi-directional LSTM-Recurrent Neural Network”. In: *Procedia Computer Science* 165 (2019). 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019, pp. 74–82. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.01.072. URL: https://www.sciencedirect.com/science/article/pii/S1877050920300806.

[4] Shlok Gilda. “Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection”. In: *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*. 2017, pp. 110–115. DOI: 10.1109/SCORED.2017.8305411.

[5] Mykhailo Granik and Volodymyr Mesyura. “Fake news detection using naive Bayes classifier”. In: *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[6] Kaggle. “Fake News Kaggle Dataset. [Online]. Available: https://www.kaggle.com/c/fake-news/data?select=train.csv”. In: (2020).

[7] Rohit Kumar Kaliyar. “Fake news detection using a deep neural network”. In: *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE. 2018, pp. 1–7.

[8] Rohit Kumar Kaliyar et al. “FNDNet–a deep convolutional neural network for fake news detection”. In: *Cognitive Systems Research* 61 (2020), pp. 32–44.

[9] Buzzfeed News. “Buzzfeed News Dataset. [Online]. Available: https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data”. In: ().

[10] Ray Oshikawa, Jing Qian, and William Yang Wang. “A survey on natural language processing for fake news detection”. In: *arXiv preprint arXiv:1811.00770* (2018).

[11] Muhammad Umer et al. “Fake news stance detection using deep learning architecture (CNN-LSTM)”. In: *IEEE Access* 8 (2020), pp. 156695–156706.