

A First Stage Project Evaluation Report on:

News Data Analysis

Prepared by :

Admission No.

Student Name

U17CO085

Kalp Panwala

U17CO104

Keshav Goyal

U17CO107

Raj Shah

U17CO113

Viren Kathiriya

Class : B.TECH. IV (Computer Engineering) 7th Semester

Year : 2020-2021

Guided by : Dr. Dipti P. Rana



**DEPARTMENT OF COMPUTER ENGINEERING
SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY,
SURAT - 395 007 (GUJARAT, INDIA)**

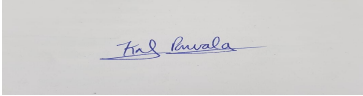
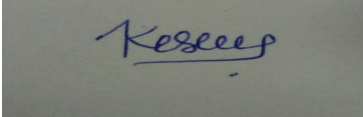
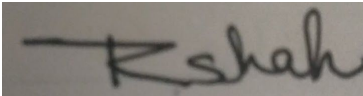

Student Declaration

This is to certify that the work described in this project report has been actually carried out and implemented by our project team consisting of

Sr.	Admission No.	Student Name
1	U17CO085	Kalp Panwala
2	U17CO104	Keshav Goyal
3	U17CO107	Raj Shah
4	U17CO113	Viren Kathiriya

Neither the source code therein nor the content of the project report has been copied or downloaded from any other source. We understand that our result grades would be revoked if later it is found to be so.

Signature of the Students:

Sr.	Student Name	Signature of the Student
1	Kalp Panwala	
2	Keshav Goyal	
3	Raj Shah	
4	Viren Kathiriya	

Certificate

*This is to certify that the project report entitled
News Data Analysis is prepared and presented by*

Sr.	Admission No.	Student Name
1	U17CO085	Kalp Panwala
2	U17CO104	Keshav Goyal
3	U17CO107	Raj Shah
4	U17CO113	Viren Kathiriya

*Final Year of Computer Engineering and their work is
satisfactory.*

SIGNATURE :

GUIDE

JURY

HEAD OF DEPT.

Abstract

The issue of online fake news has been increasing rapidly misleading many people. Because of this people finds it difficult to believe the news online. This generates the need to find new tools that can do the verification process. Thus, the goal is to find a classification model that identifies the phony features accurately using fine-grain and coarse-grain feature extraction techniques. This report proposes a framework that explores a method to combine the coarse-grain features and fine-grain features. The fine-grain features were extracted using empath library which consists of around 200 features and creates a word vector. This word vector was tested on various ML classifiers like SVM, KNN, Gradient Boost which gave an accuracy of around 90%.

Keywords Granularity - Fine Grain - Coarse Grain - Empath - Data Analysis

Table of Contents

List of Figures	vi
List of Tables	viii
List of Acronyms	ix
1 Introduction	1
1.1 Definition/Applications	1
1.2 Motivation	1
1.3 Contribution	2
1.4 Objectives	2
1.5 Organization	3
2 Literature Survey	4
2.1 Past Works/Researches	4
3 Proposed Algorithm & Implementation	10
3.1 High Level Block Diagram	10
3.2 Development of Algorithm	11
3.2.1 Dataset Description	11
3.2.2 Dataset Pre-Processing	11
3.2.3 Empath Analytics	13
3.2.4 Machine Learning Models	13
4 Simulation and Results	15
4.1 Upshots	15
4.2 Result Analysis	16
4.3 Summary	17
5 Conclusion and Future Work	18
5.1 Conclusion	18
5.2 Future Work	18

List of Figures

1. Architecture of Proposed Framework	10
2. Data Processing Techniques	13
3. News Analysis	15
4. True News World Cloud	16
5. False News World Cloud	17

List of Tables

1. Summary Of Literature Survey	7
---	---

List of Acronyms

SVM Support Vector Machine

KNN KNearest Neighbours

LDA Linear Discriminant Analysis

ANN Artificial Neural Network

NLTK Natural Language Toolkit

Chapter 1

Introduction

1.1 Definition/Applications

News varies from a simplistic review or a comment on social sites to a rumor or fake data. Its analysis and getting fruitful results can assist everyone around. E-commerce shopping and retails have become much more common now. Being profit-driven, sellers tend to spam the review to which legitimate businessmen fall in its prey. It's also true that fake news and its propagation had a non-negligible influence in politics, industries and specific markets. In this report, we experiment on the possibility to detect fake news on the basis of granularity, i.e. Coarse and fine grained features, aiming at textual data.

1.2 Motivation

The prevalence of fake news has increased with the rise of social media. Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid spread of information allow users to consume and share the news. On the other hand, it can be used to spread fake news, i.e., news with false information. The rapid spread of fake news has the potential for calamitous impacts on individuals and society. For example, the most popular fake news was more widely spread on Facebook and Twitter than the most popular authentic mainstream news during the U.S. Presidential election [16]. Therefore, fake news detection has attracted increasing attention from researchers to politicians.

Fake news detection on social media has unique characteristics and presents new chal-

lenges. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult to detect based on news content. Thus, we need to include auxiliary information, such as user social engagements on social media, to help differentiate it from the true news. Second, exploiting this auxiliary information is nontrivial in and of itself as users social engagements with fake news produce data that is big, incomplete, unstructured, and noisy. This quick guide is based on a recent survey that presents issues of fake news detection on social media, state-of-the-art research findings, datasets, and further directions.

Fake news itself is not a new problem, and the media ecology has been changing over time from newsprint to radio/television, and recently online news and social media. The impact of fake news on traditional media can be described from the perspective of psychology and social theories. For example, two major psychology factors make consumers naturally vulnerable to the fake news: (i) Naive Realism: consumers tend to believe that their perceptions of reality are the only accurate views. (ii) Confirmation Bias: consumers prefer to receive information that confirms their existing views.

1.3 Contribution

Implemented the granularity aspect wrt. Fine Grained features using empath. Data pre-processing and cleaning along with training of various models backed up with literature survey was carried out. Detailed in Chp. 3 of the report.

1.4 Objectives

Detecting dishonest behavior of retailers can make an impact and maintain the social trust on such applications. To proceed with, we used various topic models to detect such fake news on the basis of granularity. The main objective would be to divide the attributes into respective defined granularity [20] and apply Machine Learning techniques individually on them. The resultant will be combined and fed into another model and the final result gives us the statistics and analysis, beyond which we can infer further details.

1.5 Organization

This report is divided into 5 chapters. The first chapter provides the introduction to the topic. The details of fake news and techniques used in its detection. The second chapter lists out the previous work done in fake-news detection and techniques which can be employed for detection. The third chapter gives the detailed description of the framework proposed and the methods involved in it. fourth chapter gives the detailed simulation and results derived and finally fifth chapter contains conclusion and future work.

Chapter 2

Literature Survey

Fake news is intentionally written to misguide readers and leads to unexpected consequences. To detect such news, be it on social media, text contents, video streams, or image snapshots, many techniques have been researched and developed.

2.1 Past Works/Researches

Markines et al. (2009) highlighted six techniques of tagging system [1] like TagSpam, Tag-Blur, DomFp, NumAds, Plagiarism, ValidLinks. On the basis and reference of proposed features, Adaboost algorithm was implemented. They used several algorithms together and combined them using multi-voting methodologies.

To detect Deceptive reviews **Ning Cao et al. (2020)** used coarse and fine-grained features[2]. To verify the effectiveness and performance of this framework, typical LDA-BP + TextCNN model, explicit fine-grained feature mining models Unigram and POS, as well as excellent deep learning-based implicit feature mining models such as TextCNN, LSTM, and Bi-LSTM are selected and compared where LDA-BP + TextCNN model gains the best performance on the balanced/unbalanced Yelp datasets. Here LDA-BP + TextCNN was used for extracting coarse and fine-grained features while SVM was used as a classifier.

Qazvinian et al. (2011) proposed three features to identify rumors and also to identify users who believe the rumor and further spread it [3]. The three features were content-based, network-based, and twitter specific memes. For the experiment, the author collected 10,000 annotated tweets from twitter and achieved 0.95 in Mean Average Precision.

Gupta et al. (2013) analyzed the propagation of false information on Twitter regarding the 2013 Boston Marathon Bombings. To do so, they collect 7.9M unique tweets by using keywords about the event. Using real annotators, they annotate 6% of the whole corpus that represents the 20 most retweeted tweets during this crisis situation. Their analysis indicates that 29% of the tweets were false and a large number of those tweets were disseminated by reputable accounts whereas 51% were generic opinions and comments. Furthermore, they note that out of the 32K accounts that were created during the crisis period, 6 thousand of them were suspended by Twitter [4]. indicating that accounts were created for the whole purpose of disseminating false information.

Ahmed et al. (2017) proposed a fake news detection system which uses Term Frequency-Inverse Document Frequency (Tf-Idf) and n-gram analysis as feature extraction techniques. In this paper, two different feature extraction techniques and six different machine classification techniques (SVM, LSVM, KNN, DT, SGD, Logistic Regression) are used for investigating the datasets and obtain the results.

Chen, Konroy and Rubin (2015) classified the news into three types named Type A-Serious Fabrication, Type B-Large Scale Hoaxes, Type C- Humorous Fakes [6]. Serious Fabrications or Tabloids focused on sensational crime investigations , exaggeration. Type B was based on deliberate false and deceive audiences masquerade as news. Type C used to classify news on humor and identify originating sources.

Chhabra (2011) proposed a malicious website detection method which uses URL static feature based detection with accurate results. The author has taken data from a phishing platform named as 'Phishtank' which contains malicious URLs. The author has considered features such as IP addresses, a vector construction VSM[7] is chosen as the URL vector model.

Ethan Fast et al. delivered methods to classify large-scale texts into lexical categories based on semantics, emotions, reviews. Empath [8] with around 200 features which are fine grained features for given sentences. This tool was used to explore reviews on hotels, sentiment analysis of tweets, etc. This tool analytics was compared with LIWC (Linguistic Inquiry and Word Count) [14] and found to be more appealing than other tools.

P. Lakshmi Prasanna and Dr. Rao emphasized on the Artificial Neural Network [9] for text classification. They proposed this method to overcome poor results by linear statistical techniques like SVM, Regressions and initiate text classification using neural networks [10] .

ANN could also deduce unseen connections on unseen and untrained data, thus providing an unsupervised technique with proving results.

Matthew Whitehead and Larry Yaeger [13] put forth automatically classifying human sentiment from natural language written text. Bagging, Boosting [12], improvement over linear SVM were delivered. They classified reviews into negative and positive classes which further were provided as input to model for future classification.

Authors	Paper Title	Model	Dataset	Features	Advantage	Future Work
Markines et al. (2009)	Social Spam Detection using Adaboost	SVM, Ada-Boost	Spam posts on social media, tags.	TagSpam, TagBlur, DomFp, NumAds, Plagiarism, ValidLinks	Use of Boosting is better than bagging to classify.	Include multiclass classification rather than binary.
Ning Cao et al. (2020)	A deceptive review detection framework	LDA-BP + TextCNN + SVM	Yelp Datasets	Fine-grained and coarse-grained features	Proposed model out-performed TextCNN, LSTM, and BiLSTM	Improvement Possible by integrating other better deep neural network algorithms and coarse-grained algorithms
Qazinian et al. (2011)	Identified tweets in which rumor is endorsed.	Naive Bayes	Tweets	Content-based, network-based, Twitter-specific memes	Rumor analysis with 0.95 Mean Average Precision	To build a system to identify whether a trending topic is a rumor or not
Gupta et al. (2013)	Analysis of Twitter content during Boston Marathon.	Logistic Regression	Tweets and user information	Topic engagement, Global engagement, Social reputation, Likability, Credibility	To find out fake Twitter accounts created amid high impact news events	Improvement possible by using a decision tree can add culture affected feature.

Ahmed et al. (2017)	Detection of Online Fake News Using N-Gram Analysis and ML Techniques	SVM, LSVM, KNN,DT, SGD, Logistic Regression	News articles (Reuters), Fake news dataset (kaggle)	TF-IDF	TF-IDF out-performed n-gram analysis	Include more feature extraction techniques
Chen, Konroy and Rubin	Deception Detection and 3 Types of Fakes	NLP, Sentiment Analysis, Big Data	E-Mails, Web Crawling, Fake product reviews, publicly available social media data, Law enforcement data	Type A - Serious Fabrication, Type B - Large Scale Hoaxes, Type C - Humorous Fakes	Broader analysis of news data along with fake classification	Classify the divisions further based on cultures and religions
S. Chhabra et al. (2011)	Fake and Malicious URL Detection	Naive Bayes, Logistic Regression, DT, SVM-RBF, SVM-Linear SVM-Sigmoid	Malicious URL dataset from 'Phishtank'	Grammar, Lexical, Vectors and Static.	Proposed URL static feature based detection method	Can use binary classification through Adaboost algorithm.

Ethan Fast, BinBinbin Chen, Michael Ber-stein	Empath: Under- standing Topic Signals in Large- Scale Text	Empath	Hotel Reviews, Movie Reviews, time analysis of mood on Twitter, deception data	Text classification, neural network training, 200 in-built features	Simplified fine-grained classification	More qualitative aspect of classification, include fiction reviews.
P. Lakshmi Prasanna, Dr. Rao	Text classification using artificial neural networks	ANN, Document conversion, stemming	Multi- dimensional datasets From medical background	TFIDF Matrix from text classification.	ANN is better than linear statistical techniques	Implement NN on text.
Matthew Whitehead, Larry Yaeger	Sentiment Mining Using Ensemble Classification Models	Bagging, Boosting [12], single model SVM, K-fold(10) cross validation	Dataset by Snyder and Barzilay, Amazon, layerratingz, tvratingz	Ensemble models give much better accuracy	Out- performed SVM model	Improvement in time complexity

Table 2.1: Summary Of Literature Survey

Chapter 3

Proposed Algorithm & Implementation

Coarse grained features are explicitly defined as overall data in the text which has a tendency to split enough. The smallest possible meaningful content in a topic model can be a word which defines Fine Grained features. When combined together define the coarse grained features being a superset. High level block diagram is explained thoroughly in further sections.

3.1 High Level Block Diagram

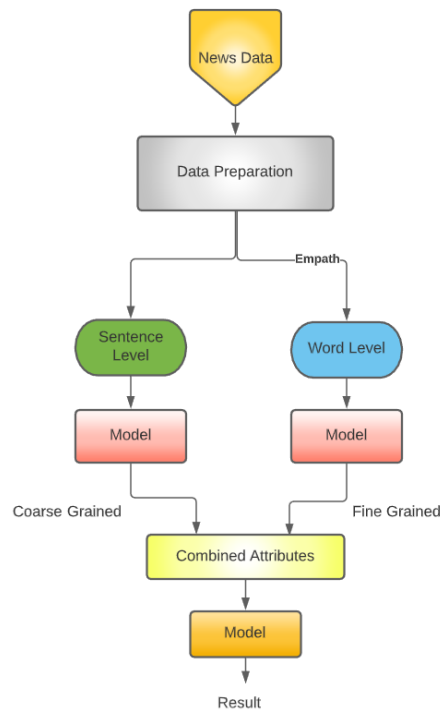


Fig 1. Architecture of Proposed Framework

The dataset is initially processed as described in above sections. Classification based on granularity is imposed and features are extracted on the basis of proposed algorithms. Fine Grained features are retracted using **Empath** library of Python. Further modifications of dataset is carried out. Training the dataset on models detailed above is carried out. Results are carried out and combined with coarse grained features. Thus, word level and sentence level attributes play a vital role to analyse the news data. The resultant vector of attributes are trained using Supervised ML technique to further emphasize on best analysis so as to avoid deceptions.

3.2 Development of Algorithm

3.2.1 Dataset Description

This dataset consists of about **8000** articles consisting of fake as well as real news. Our aim is to train our model so that it can correctly predict whether a given piece of news is real or fake. The fake and real news data is given in two separate datasets with each dataset consisting around 4000 articles each. The features of the dataset are title, news subject, date, text.

3.2.2 Dataset Pre-Processing

Data preprocessing [17] is one of the most required steps in data analysis in order to achieve maximum accuracy and throughput . It includes techniques to remove incomplete data, making data consistent and ready to use for experiments. Mostly, library called pandas [19] is used for such preprocessing.

Preprocessing text simply means to convert text into a form that is predictable and analyzable for given task. Various steps involved in Data Preprocessing are Data Cleaning, Data Transformation, Data Reduction.

Data Cleaning involves handling of missing data, noisy data. Strategies to handle missing data involve removing the tuples, filling the missing values. In noisy data Lowercasing, Stemming, Lemmatization, Stop words removal, outlier analysis can be done to clean irregular and inconsistent data.

Lowercasing is one of the simplest and most effective form of text preprocessing. **Stemming** is the process of reducing inflection in words (e.g. changing, changed, changer) to their root form (e.g. change). The "root" in this case may not be a real root word, but just a canonical

form of the original word. **NLTK** provides implementation of stemming. Stemming is desirable as it may reduce redundancy as most of the time the word stem and their derived words mean the same.

Lemmatization is very similar to stemming, where the main aim is to remove inflections and map a word to its root form. The only difference is that, lemmatization tries to do it properly. It doesn't just remove things off, it links words with similar meaning to one word.

Stop words are a set of commonly used words in a language. Examples of stop words in English are articles, etc. The intuition behind removing stop words is that, we can focus more on the important words instead.

Noise removal is about unnecessary punctuations and white spaces that can interfere with your text analysis.

Data Transformation is used to transform the data in appropriate form which makes it usable for data mining process. There are different ways to transform your text also for each task different methods are followed.

Text Normalization is the process of transforming a text into a canonical (standard) form. For example, the "3" can be transformed to "three", its canonical form. It is a highly overlooked preprocessing step.

Data Reduction is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis becomes harder. In order to get rid of this, we use techniques like Dimensionality Reduction, Feature Selection, etc. It aims to increase the throughput, efficiency and reduce data storage and analysis costs.

Existing rows that contain irregular and incomplete data were identified and removed, so as to manipulate the dataset [21]. The rows of the dataset were shuffled. The dataset was further divided into 70% training data and 30% validation data.

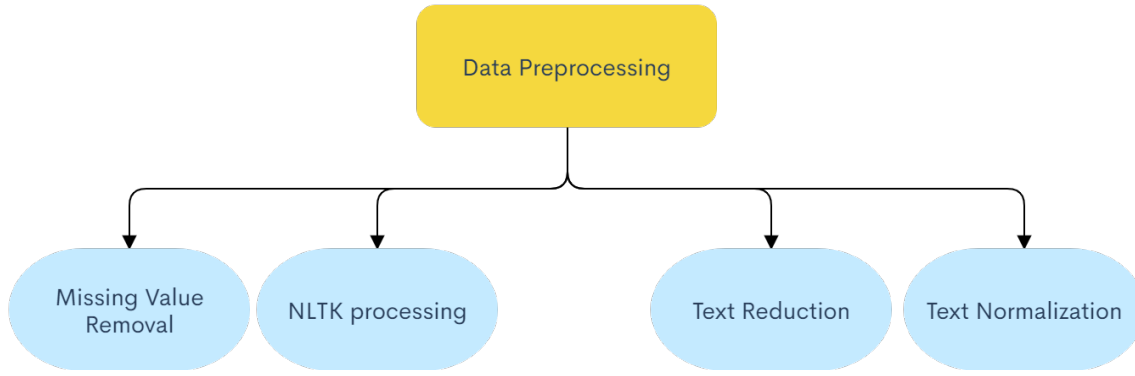


Fig 2. Data Processing Techniques

3.2.3 Empath Analytics

Empath is a tool for analyzing text across lexical categories, and also generating new lexical categories to use for an analysis. It contains 194 inbuilt attributes, semantic meaning of text can easily be represented in the form empath feature vector.

These attributes represent the columns of the modified dataset with rows as the empath values of the existing data. A new attribute named Category is appended to the modified dataset which can further be useful as an input to supervised machine learning techniques.

3.2.4 Machine Learning Models

Various supervised machine learning models were implemented on the dataset to bridge the gap between fake and real news.

Different models such as Logistic Regression, Linear Support Vector Machine, Gradient Boosting, K-Nearest Neighbour, Naive Bayes and tree algorithms(Random Forest, Decision Tree) [18] were implemented and analyzed. Confusion matrix was built for each of the model and best suitable algorithm was found to be random forest. The accuracy [15] touched to an extreme of 90% on the testing dataset. Thus, the fine grained features would be further shuffled with coarse grained analysis resulting in better prediction on news data.

Algorithm 1: Fine-grained feature acquisition

Input: *Dataset*

Output: *Fine-grained feature*

Process:

- 1: *Prepare the dataset.*
 - 2: *For $t = 1$ to n : // n is the number of the inbuilt empath attributes:*
 - 3: *Apply empath attributes on textual data.*
 - 4: *Modify the dataset using these attributes by taking a transpose.*
 - 5: *Apply Supervised Machine Learning Algorithm.*
 - 6: *Get the fine-grained feature vector and corresponding results.*
 - 7: *End for.*
-

Chapter 4

Simulation and Results

4.1 Upshots

On the basis of previous discussions, implementation and its analysis was carried out on the basis of news granularity. Fine Grained descriptions are put forth with Coarse Grained yet to be displayed.

The analysis of type of news and its subject is displayed below. Further, they are classified into its corresponding fake and true classes. The news varied from Politics, world news to government, Middle-East, US election news.

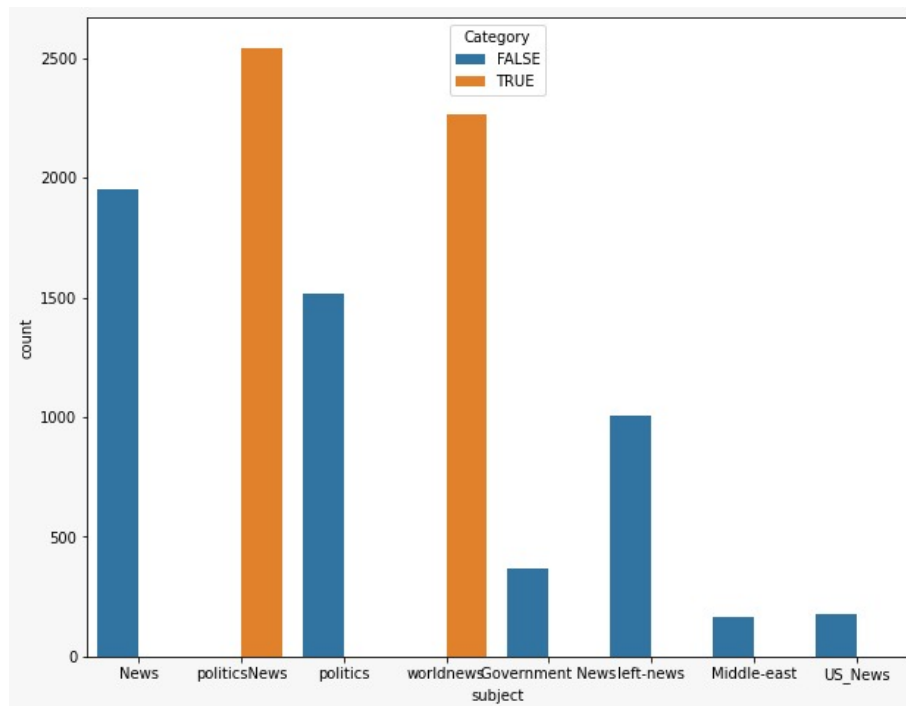


Fig 3. News Analysis

Chapter 5

Conclusion and Future Work

5.1 Conclusion

We conclude a part of granularity , ie. Fine Grained on textual news. Data Preprocessing was a core part along with selection of Empath library that drives deeper into word level features for analysis of the required news.

5.2 Future Work

Analyse sentence level features and implement model that merges with results of Fine Grained and Coarse Grained features to predict the fakeness of news appropriately so as to prevent any loss due to false predictions by model, if any.. Further, going into level of analysis where one can predict fake news on basis of caste, religions is future aim of research.

References

1. B. Markines, C. Cattuto, F. Menczer, in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (ACM, 2009), pp. 4148.
2. Ning Cao, Shujuan Ji, Dickson K.W. Chiu, Mingxiang He, Xiaohong Sun, supported in part by the Natural Science Foundation of China (No. 71772107, 61502281)
3. V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, 2011), pp. 1589-1599
4. Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. \$1.00 per rtboston marathon prayforboston: Analyzing fake content on twitter.
5. Ahmed, Hadeer Saad, Sherif. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. 127-138. 10.1007/978-3-319-69155-8_9.
6. ASIST 2015, November 6-10, 2015, St. Louis, MO, USA. Copyright © 2015 Yimin Chen, Victoria L. Rubin and Niall J. Conroy.
7. S. Chhabra, A. Aggarwal, F. Benevenuto, P. Kumaraguru, in Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (ACM, 2011), pp. 92-101.
8. Empath: Understanding Topic Signals in Large-Scale Text ACM Classification Keywords H.5.2. Information Interfaces and Presentation: Group and Organization Interfaces.
9. Text classification using artificial neural networks: International Journal of Engineering Technology Website: www.sciencepubco.com/index.php/IJET
10. Chaitanya Naik, Vallari Kothari, Zankhana Rana, Document Classification using Neural Networks Based on Words, In: International Journal of Advanced Research in Computer Science, 2015.
11. Snyder, B., and Barzilay, R. 2007. Multiple aspect ranking using the good grief algorithm. In Proceedings of NAACL HLT, 300307.

12. Schapire, R. E. 2002. The boosting approach to machine learning: An overview
13. Sentiment Mining Using Ensemble Classification Models Conference Paper, January 2008
14. James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. In Mahway: Lawrence Erlbaum Associates 71 2001.
15. Confusion Matrix-based Feature Selection :Conference Paper. January 2011 by Sofia et al.
16. Amador Diaz Lopez, J., Oehmichen, A., Molina-Solana, M.: Fakenews on 2016 US elections viral tweets (November 2016March 2017), November 2017.
17. Review of Data Preprocessing Techniques in Data Mining : Article in Journal of Engineering and Applied Sciences. September 2017 DOI: 10.3923/jeasci.2017.4102.4107
18. A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM : Jour of Adv Research in Dynamical Control Systems, Vol. 11, 06-Special Issue, 2019
19. W. McKinney, pandas: a python data analysis library, [http: //pandas.sourceforge.net](http://pandas.sourceforge.net)
20. Fake News Detection Method Based on Text-Features: IMMM 2019 : The Ninth International Conference on Advances in Information Mining and Management
21. Research on Data Preprocessing in Exam Analysis System : College of Computer and Information Engineering, Jiangxi Normal University, Nanchang330022, china

Acknowledgements

At the outset, we thank the God Almighty for the grace, strength and hope to make our endeavor till now a success. We express our sincere gratitude to our guides Dr. Dipti P Rana, Assistant Professor, Computer Engineering Department, SVNIT and Ms. Isha Agarwal, PhD Scholar, Computer Engineering Department, SVNIT for their exemplary guidance, monitoring and constant encouragement without which this report would not have been possible.

We would like to thank Dr. Mukesh A. Zaveri, Head of Dept. Computer Engineering, SVNIT for support provided.

We would like to thank all who directly or indirectly contributed to this report.