News Summarization and Authentication

Mrunal Waydande, Vidhu Priya, Bhawana Thavarani, Prof. Dr. Dipashree Bhalerao

Sinhgad College of Engineering, Dept. of Electronics and Telecommunications

---***---

Abstract - The proposed system supports two major functionalities, namely authentication and summarization. The module of authenticity is given more priority over the summarization; therefore, authentication of news leads to summarization of news such that if authenticity measure is greater than or equal to 55% only then the summarization module will execute or else system will return the authenticity parameters of news in terms of percentage. The problem of fake news must be addressed using artificial intelligence approaches and machine learning algorithm, Syntax net algorithm is used to extract phrases from user input Statistical analysis and keywords have been used in this system for verification of news and ensuring whether the news article is a hoax news by comparing the contents from some allowed news website with the input news.

***keywords*:** Artificial Intelligence, Machine Learning, Naïve Bayes classifier, News, Support Vector Machine (SVM), Text Rank, TF-IDF

1. INTRODUCTION

This In Today's world, anybody can post the content over the internet. Unfortunately, counterfeit news gathers a lot of consideration over the web, particularly via web-based networking media. Individuals get misdirected and don't reconsider before flowing such mis-educational pieces to the most distant part of the arrangement. Such type of activities are not good for the society where some rumors or vague news evaporates the negative thought among the people or specific category of people. As fast the technology is moving, at the same pace, the preventive measures are required to deal with such activities. Broad communications assuming a gigantic job in affecting the public and, as it is normal, a few people attempt to exploit it. There are many sites which give false data. They deliberately attempt to bring out purposeful publicity, deceptions and falsehood under the pretense of being genuine news. Their basic role is to control the data that can cause open to have confidence in it. There are loads of case of such sites everywhere throughout the world. Therefore, counterfeit news influences the brains of the individuals. As showed by study, Scientist accept that many man-made brainpower calculations can help in uncovering the bogus news. Fake news detection is made to stop the rumors that are being spread through the various platforms whether it be social media or messaging platforms, this is done to stop spreading fake news which leads to activities like mob lynching, this has been a great reason motivating us to work on this project. We have been continuously seeing various news of mob lynching that lead to the murder of an individual; fake news detection works to detect this fake news and stopping activities like this protecting the society from these unwanted acts of violence.

The primary aim is to detect the fake news, which is a classic text classification problem with a straightforward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news. This leads to consequences in social networking sites like Facebook, Instagram, microblogging sites like Twitter and instant messaging applications like WhatsApp, Hike where these fake news gets a major boost and gets viral among people, around the country and globe. The proposed system helps to find the authenticity of the news. If the news is not real, then the user is suggested with the relevant news article. Shows the suggested format and appearance of a manuscript prepared for SPIE journals. Accepted papers will be professionally typeset. This template is a tool to improve manuscript clarity for the reviewers. The final layout of the typeset paper will not match this template layout.

2. LITERATURE SURVEY

In (Improving spam detection in Online Social Networks 2015), author proposed how the massive use online social media has increased nowadays, with great advancement comes great responsibility. Similarly, author classified users into two categories, namely spammers and non-spammers. The primary aim of this system was to detect spammers in twitter Online social network. This is a very important aspect for our project! Techniques used to detect the spammer account present on a social media platform with accuracy 87.9% are: first, the data is classified as spammer and non-spammer using a learning algorithm such as.

I) Naive Bayes: given the attributes of the user account, the probability associated with the user of being a spammer was generated.

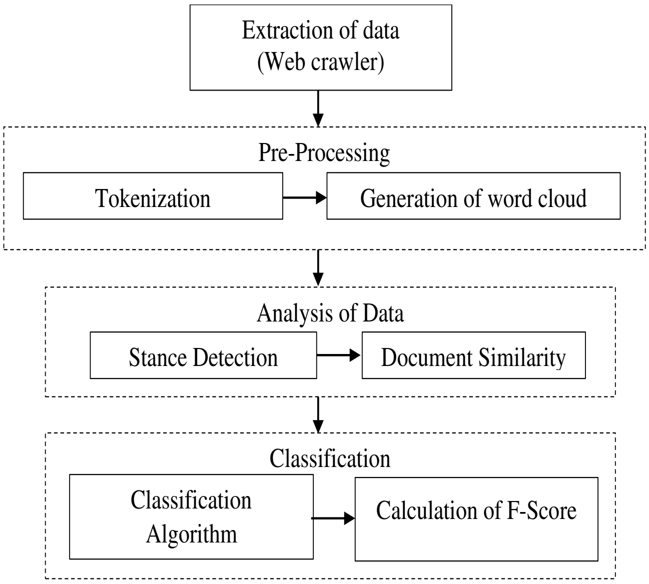
II) Clustering was unsupervised learning algorithm which can classify a given user into spammer/non-spammer.

III) Decision tree was also used for classification of user account. In this method, using the decision tree, a decision was made at every level of the decision tree.

In (automatic text summarization: a reader-oriented approach 1994), author proposed a new method for text summarization by specifying the informative and condensed nature of the input, which is a very important aspect for our system. The method used in this system is called ROSE- Reader Oriented Summary Engine; the primary goal of this engine was to generate the summary from the perspective of user rather than generating summary from the perspective of text. In (Spotting Fake News: A Social Argumentation Framework for Scrutinizing Alternative Facts) author proposes a prototype to verify the proposed alternative fact to reduce the proliferation of fake news by describing the impact of fake news over the internet. Measures were taken in order to help users to verify the alternative facts, most of the time fake news emerges through improper knowledge of subject base. Every other system verifies the news with some ground knowledge of truth, whereas this system matched the most relevant news with the prescribed news detail. Approaches such as analyzing the conclusion, the quality of arguments are judged, and deciding portions of issue to be analyzed. This system not only could detect the fake,

News also provides the statistical analysis to user to analyze future fake news article. In, (Challenges in automatic summarization) author provides an insight into all the problems involved in designing an effective text summarization technique. With development alongside the researchers are investigating and implementing summarization tools to extract meaningful information from the natural language input to a machine. The nature of the summary can differ from one algorithm to another. The reason for this difference is associated in the ground level, whether they are extractive or abstractive.

3. BLOCK DIAGRAM



4. IMPLEMENTATION

➢ AUTHENTICATION

Naive Bayes

A Naive Bayes classifier is a supervised machine learning algorithm that uses Bayes‟ theorem. The variables that are used to generate the model are independent of each other. It is proven that this classifier itself provides pretty excellent results. The classification is conducted by deriving the maximum posterior, which is the maximal P (Ci|X) with the above assumption applying to Bayes' theorem. This assumption reduces the computational cost by only counting the class distribution. Naive Bayes is a popular algorithm which is used to find the accuracy of the news, whether it's real or fake using multinomial Naïve Bayes. There are several algorithms that focus on common principle, so it is not the only algorithm for training such classifiers. To check if the news is fake or real, naïve Bayes can be used.

Support Vector Machine (SVM)

SVM is a good algorithm to extract the binary class based on the data given to the model. In the proposed model, the work is to classify the article in two categories either true or false. A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both regression and classification purposes. It is based on the idea of finding the hyper-plane that best divides the dataset into two classes. Hyper-planes are decision boundaries that help the machine learning model classify the data or data points. How the classification of the data.

➢ SUMMARIZATION

Text Rank

TextRank is a graph-based ranking algorithm under the hood for ranking chunks of text segments in order of their importance in the text document. In order to find the most relevant sentences in text, a graph is constructed where the vertices of the graph represent each sentence in a document and the edges between sentences are based on content overlap, namely by calculating the number of words that 2 sentences have in common. Based on this network of sentences, the sentences are fed into the PageRank algorithm which identifies the most important sentences. When we want to extract a summary of the text, we can now take only the most important sentences. In order to find relevant keywords, the textrank algorithm constructs a word network. This network is constructed by looking which words follow one another. A link is set up between two words if they follow one another, the link gets a higher weight if these 2 words occur more frequently next to each other in the text. On top of the resulting network the PageRank algorithm is applied to get the importance of each word. The top 1/3 of all these words are kept and are considered relevant.

After this, a keywords table is constructed by combining the relevant words together if they appear following one another in the text.

TF-IDF

TFIDF, short for term frequency–inverse document frequency, is a numeric measure that is use to score the importance of a word in a document based on how often did it appear in that document and a given collection of documents. The intuition behind this measure is: If a word appears frequently in a document, then it should be important and we should give that word a high score. But if a word appears in too many other documents, it’s probably not a unique identifier, therefore we should assign a lower score to that word.

Formula for calculating tf and idf:

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

IDF(w) = log_e(Total number of documents / Number of documents with term w in it)

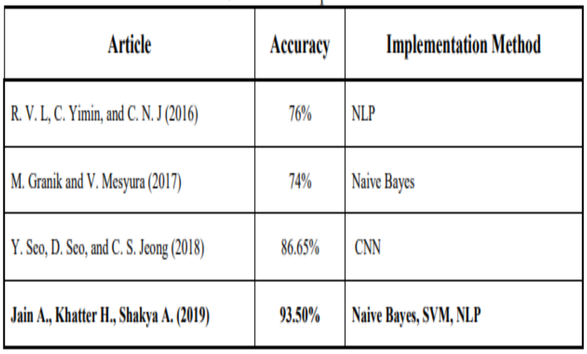
Hence tfidf for a word can be calculated as:
TFIDF(w) = TF(w) * IDF(w)
Word Frequency

Word Frequency is an algorithm that counts how many times a word appears in a document. It's a tally. Those word counts allow us to compare documents and gauge their similarities for applications like search, document classification and topic modeling. Word Frequency is a also method for preparing text for input in a deep-learning net. Word Frequency lists words paired with their word counts per document. In the table where the words and documents that effectively become vectors are stored, each row is a word, each column is a document, and each cell is a word count. Each of the documents in the corpus is represented by columns of equal length. Those are wordcount vectors, an output stripped of context.

5. RESULTS

The results of mentioned four models are compared with the proposed model, it is found the accuracy among top 200 results is mentioned in the table. The demonstration is done using python programming on R studio and some machine learning algorithm

Table 5.1 Result Comparison



Article	Accuracy	Implementation Method
R. V. L, C. Yimin, and C. N. J (2016)	76%	NLP
M. Granik and V. Mesyura (2017)	74%	Naive Bayes
Y. Seo, D. Seo, and C. S. Jeong (2018)	86.65%	CNN
Jain A., Khatter H., Shakya A. (2019)	**93.50%**	**Naive Bayes, SVM, NLP**

6. CONCLUSIONS

It is significant to find the accuracy of news which is available on internet. In the paper, the components for recognizing Fake news are discussed. A mindfulness that not all, the fake news will propagate via web-based networking media. Currently, to test out the proposed method of Naïve Bayes classifier, SVM, and NLP are used. In the future, ensuing algorithm may provide better results with hybrid approaches for the same purpose fulfilment. The mentioned system detects the fake news on the based on the models applied. Also, it had provided some suggested news on that topic which is very useful for any user. In the future, the efficiency and accuracy of the prototype can be enhanced to a certain level, and also enhance the user interface of the proposed model.

REFERENCES

- M. Granik and V. Mesyura, "Fake news detection using naïve Bayes classifier," 2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc., pp. 900–903, 2017.
- https://indianexpress.com/article/technology/social/whatsapp-fight-against-fake-news-topfeatures-to-curb-spread-of-misinformation5256782/Martínez-Garcia, S. Morris, M. Tscholl, F. Tracy, and P. Carmichael, "Case-based
- learning, pedagogical innovation, and semantic web technologies," IEEE Trans. Learn. Technol., vol. 5, no. 2, pp. 104–116, 2012
- S. Gilda, "Evaluating machine learning algorithms for fake news detection," IEEE Student Conf. Res. Dev. Inspiring Technol. Humanit. SCOReD 2017 - Proc., vol. 2018–January, pp. 110–115, 2018.
- Y. Seo, D. Seo, and C. S. Jeong, "FaNDeR: Fake News Detection Model Using Media Reliability," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2018– October, no. October, pp. 1834–1838, 2019.
- S. Das Bhattacharjee, A. Talukder, and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018–January, pp. 556–565, 2018.
- S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news etection on Twitter," Proc. 2018 IEEE/ACM Int.Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018, pp. 274–277, 2018