**Fake News Detection**

Naimul Hasan Akib. Nitu khatun

Department of Computer Science Department of Computer Science

and Engineering and Engineering

Daffodil International Daffodil International

University University

Dhaka, Bangladesh. Dhaka, Bangladesh

[Naimul15-2423@diu.edu.bd](mailto:Naimul15-2423@diu.edu.bd) Nitu15-2747@diu.edu.bd

Kalpana Akter. Afzazul Haque Aungkur

Department of Computer Science Department of Computer Science

and Engineering and Engineering

Daffodil International Daffodil International

University University

Dhaka, Bangladesh. Dhaka, Bangladesh

[Kalpana15-2597@diu.edu.bd](mailto:Kalpana15-2597@diu.edu.bd) [afzajul15-2303@diu.edu.bd](mailto:afzajul15-2303@diu.edu.bd)

**Abstract**

Fake news is news designed to deliberately spread hoaxes and disinformation. It denotes type of yellow journalism which intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing. This an evolution in human history but at the same time it unfocused the line between true media and meliciously forged media.today anyone can publish content credible or not that can be consumed by the world web.sadly fake news accumulates a great deal of attention over the internet.specially on social media.

**Introduction**

Fake news refers to information content that is false, misleading or whose source cannot be verified. This content may be generated to intentionally damage reputations, deceive, or to gain attention. Social media platforms are incredibly influential. In today’s world, it is normal to receive news from online sources like social media. News is often subjective to readers. We often choose to ingest content that appeals to the different emotions we have. So, considering this, the information that gets the most reach may not be real or accurate news. Additionally, real news may be twisted in transmission. A reader may end up with different versions of the same news. This may lead to information overload.At a time when the globe is defined by a pandemic, public health depends on reliable information. Yet we stare down the barrel of an infodemic. An infodemic is the combination of the word information and epidemic. It is an excessive amount of information about a problem that makes the solution more difficult. It also defines a wide and rapid spread of misinformation.

**Backgrounds study /Literature Review**

The research questions were developed with the purpose of presenting an overview of the area, highlighting key aspects of the primary studies.

Controlled Experiment is a form of experimental study in which the researcher has control over the main aspects of the study and the independent variables being studied. Its objective is to confirm theories, conventional knowledge, to evaluate the prediction of models or to validate measurements. It involves the formulation of hypotheses to be verified in relation to the results obtained (Wohlin et al., 2012).

According to Yin (2001), the Case Study is a research strategy that encompasses a method that encompasses everything in specific approaches to data collection and analysis. It is useful when the phenomenon to be studied is broad and complex and cannot be studied outside the context in which it occurs naturally.

As a guarantee of research questions quality, the PICO (Population, Intervention, Control and Outcomes) model was used (James et al., 2016), whose objective is to prove the ability to characterization and classification of the issues. With this model, it is also possible to evaluate the effects of an intervention on a given population.

**Methodology**

A machine learning approach uses machine learning algorithms to detect misinformation. Examples of these algorithms include:

**Naïve Bayes:** uses probabilistic approaches based on Bayes theorem. This algorithm is often used for text classification.

**Decision Tree:** a supervised learning algorithm that has a tree-like flow. It helps in decision making. A useful algorithm for both classification and regression tasks.

**Random forest:** simply a combination of decision trees.

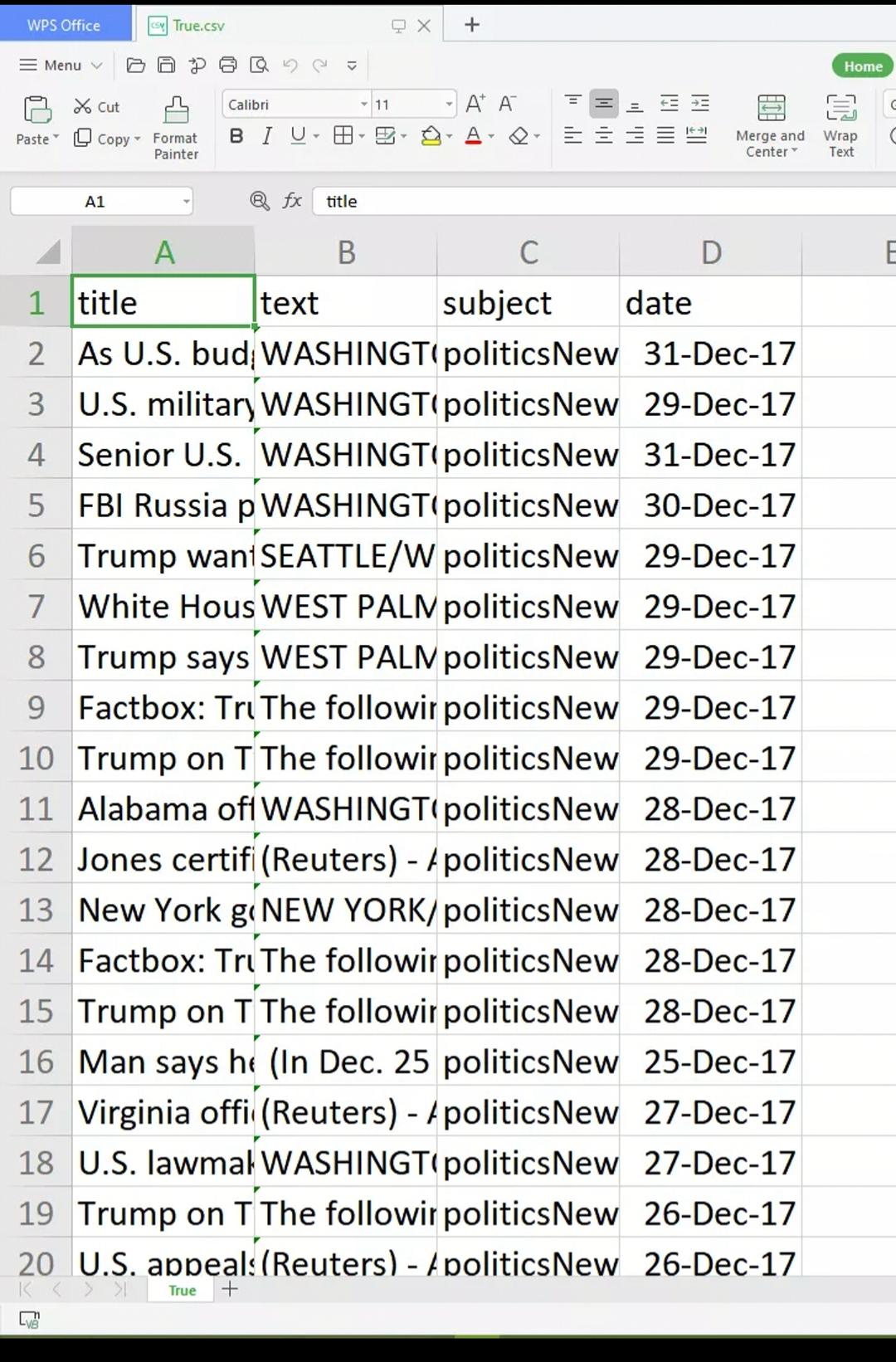
**Support Vector Machine:** a supervised learning algorithm. It examines data for classification and regression analysis. It classifies data into two categories.

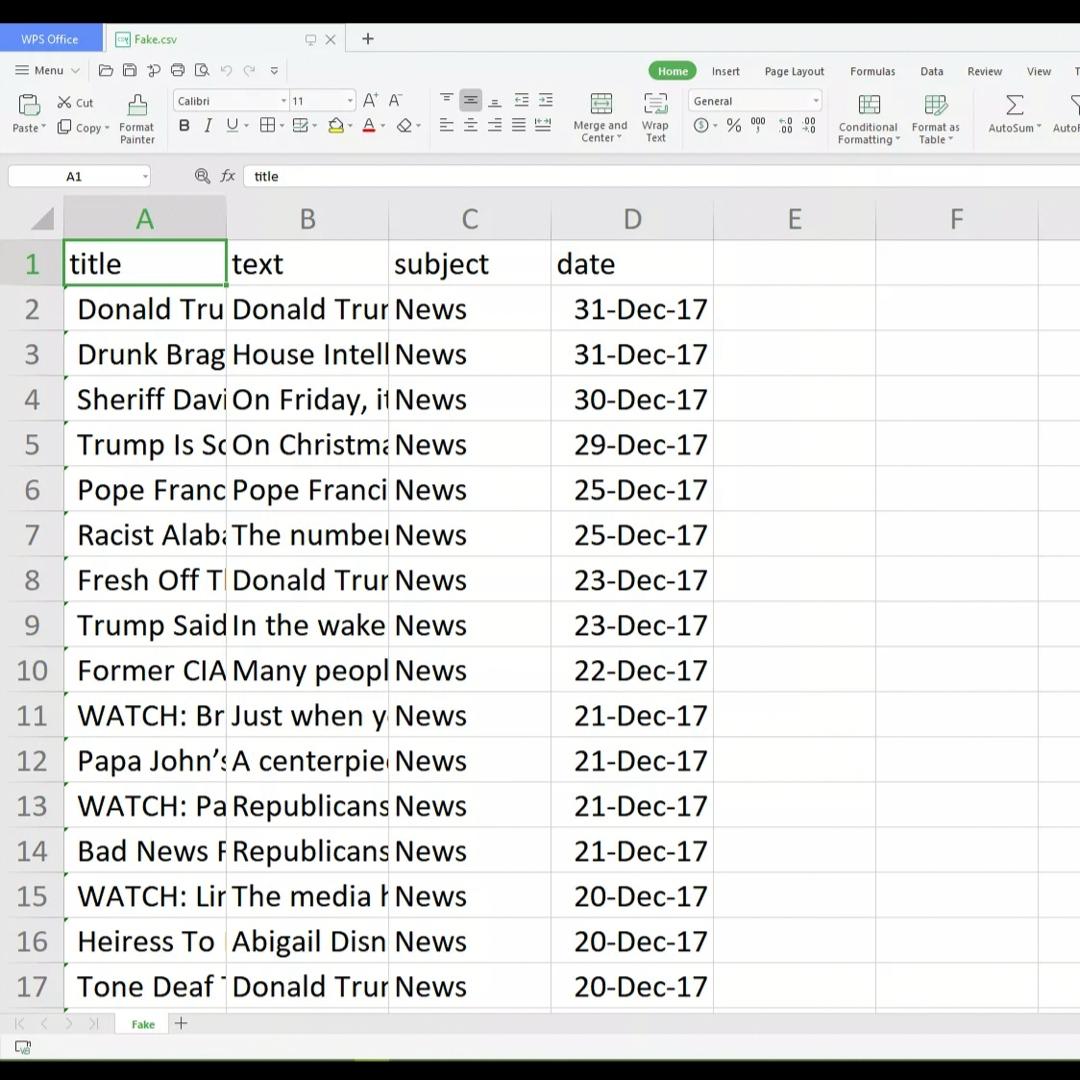
**Logistic Regression:** contrary to the name, it is a classification algorithm used to estimate discrete values.

**K-nearest-neighbor:** a simple algorithm that is used for both classification and regression tasks. Though it is more widely used for classification problems.

Datasets are used to refine the algorithms. These datasets may be split as training data or test data. I have come across a lot of research where a system combines various machine learning algorithms and data mining. This is often carried out on social media platforms, especially Twitter data. Depending on the nature of the data, the two classifiers can be applied to a dataset and their performance compared. On the other hand, these classifiers can be used in an ensemble method to enhance each other’s results in classification tasks, therefore improving model accuracy. As mentioned above, Naïve Bayes is popular in text classification tasks therefore it’s considered for such tasks often.A methodology may involve building classifiers to predict the validity of news based only on news content. This may be achieved using Recurrent Neural Network (RNN) models and long-short term memories (LSTM).

A methodology may involve building classifiers to predict the validity of news based only on news content. This may be achieved using Recurrent Neural Network (RNN) models and long-short term memories (LSTM).





**Result and Analysis**

The accuracy of the results is usually determined by the combinations of models used and the datasets involved. A combination of available toolkits with Bayesian learning may be used to develop a fake news detector. These toolkits include Textblob, Natural Language, and SciPy. But a challenge exists with some of these traditional machine learning approaches. They treat fake news detection as a binary classification task. These models alone struggle to contextualize text data. They need structured/labeled data. In fact, machine learning models struggle to solve complex queries with huge amounts of data. This is where deep learning models come into play. Logistic regression is a relatively simpler model but achieved an average accuracy of over 90% on the three datasets. There can be multiple explanations for achieving a high average accuracy; firstly, logistic regression model is fine-tuned using an extensive grid search with different hyperparameters; secondly, some of the datasets have similar writing styles of authors, which led to 97% accuracy of logistic regression model.

**Conclusion**

We discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We used multiple performance metrics to compare the results for each algorithm. The ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

**Reference**

**1.**J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.

View at: Google Scholar

**2.** A. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016.

**3.** J. Hua and R. Shaw, “Corona virus (covid-19) “infodemic” and emerging issues through a data lens: the case of China,” International Journal of Environmental Research and Public Health, vol. 17, no. 7, p. 2309, 2020.

View at: Publisher Site | Google Scholar

**4.** K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

View at: Publisher Site | Google Scholar

**5.** W. Y. Wang, Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.

**6.** N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

**7.** T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class adaboost,” Statistics and its Interface, vol. 2, no. 3, pp. 349–360, 2009.

View at: Publisher Site | Google Scholar.