ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Detecting Fake News Articles Using NLP (Natural Language Processing) to Identify In-Article Attribution as a Supervised Learning Estimator

Thallapally Sai Sankeerth

Department of Computing Technologies, SRM Institute of Science and Technology, Chennai,India.

Ashwin Sukumar

Department of Computing Technologies, SRM Institute of Science and Technology, Chennai,India.

Dr. C. Jothi Kumar

Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Chennai,India

Abstract - Intentionally false material masquerading as credible journalism is a global issue of accuracy and integrity that impacts people's opinions, decisions, and voting habits. The majority of so-called "fake news" is first disseminated via social media platforms like Facebook and Twitter before becoming mainstream media outlets like conventional television and radio news. Fake news articles published initially over social media platforms contain crucial language traits, such as overuse of unjustified exaggeration and unattributed quoted information. Currently, everyone relies on a range of online news sources since the internet is ubiquitous everywhere. Facebook, Twitter, and other social media platforms allow news to move swiftly among a large number of individuals in a short period. As such, false information can increasingly circulate over a short period, causing massive harm to many individuals. The outcomes of false news detection are examined in this research. Analysis of the news stories using natural language processing and the Naive Bayes classifier or algorithm and the SciPy tools is used to determine whether a story is authentic or fraudulent.

INTRODUCTION

With the rise of social media sites like Facebook and Twitter, most people's lives are now dominated by online activities such as utilizing these platforms regularly. Most false news currently circulates on social media platforms rather than publications and then spreads to other channels. Finding out where and how a news story originated is one of the most challenging aspects of using social media, which travels quickly and is tough to track down. As a result, most people believe that the news is genuine rather than phony since it has been widely disseminated. All members of society will feel a negative influence due to the dissemination of false information. Social media platforms have had a low bar when it comes to reporting false news since they are the only platform that has the power to fast disseminate information worldwide in a matter of seconds.

Fake news has become an issue for everyone, not only in the past but now today. As a result, it has significantly impacted many people and organizations. It would not be possible to tell which news was phony and authentic in the old days. The only thing achieved was to make it more difficult for consumers to determine whether or not a story is genuine or fake. Even though no one knows about the information in the news, people tend to accept it as fact in most circumstances [5]. In the wake of the introduction of new technology, fake news detection has been a lot easier in recent years because of the rapid advancement of technology. However, with the invention of the NLP (Natural Language Processing) software, it is now possible to process the text by classifying it based on the different algorithms to detect its credibility [6]. Based on probabilistic word counts, such as verbs and unique words in a text, the credibility of the news is measured. The Multinomial Naive Bayes classifier generates the probability score and detects false and authentic communication in this research.

LITERATURE REVIEW

Fake news has always been a source of concern for the general public, but it is much more so now. Regardless of whether you are a person or a business, it has significantly impacted your life. Initially, it was not possible to tell the difference between real and phony news back in the old days. The only achieved objective is to make it more difficult for consumers to determine whether or not a story is authentic or a hoax. When people get the news, they have no idea what the story is, therefore, fake. In other words, as time went on, technology became a reality, and detecting false news has also gotten a lot easier because of new rising trends in technology.

Detection of fake news is possible in a variety of methods. One technique to detect and debunk false information is to check the facts. Fake news often contains grammatical errors. The news will have its material highlighted in a variety of ways. To influence the reader's perceptions, they often attempt to draw attention to certain parts of the text or news [4]. Also, the text's source is not

accurate. There will be a variety of URLs to choose from, such as lower-case aplhabets, upper-case alphabets, or a mix of both, as well as symbols and digits in the URL. Another tactic the attackers use is to click on buttons that have been highlighted and are thus more noticeable to readers or visitors.

Several sophisticated Python packages for natural language processing jobs may be used. NLP library spaCy, which includes pre-trained models and support for tokenization and training in more than 60 languages, is one of the most popular [1]. The robustness and production-readiness of spaCy's software make it ideal for usage in commercial goods. The Greek Fake News Detector app developed using this library is considered more accurate and efficient. Additionally, Streamlight is a Python framework that enables the rapid development of web applications in data research [1]. In only a few lines of code, you can quickly construct a user interface using a variety of widgets. It is also an excellent tool for delivering machine learning models to the web, as well as for creating beautiful visualizations of your data. Streamlight Additionally, Streamlight features a robust caching mechanism that significantly enhances the speed of your app [1]. The library's designers provide the free Streamlit Sharing service, making it simple to distribute and share your program with others.

METHODOLOGY

The methodology applied in this research comprises using the Multinomial Naïve Bayes algorithm proposed with the help of python programming language. Python is a free and open-source integrated development environment (IDE). Machine learning can be quickly developed using Python's vast extensions and libraries. When it comes to helping computer systems understand human language, natural language processing (NLP) is the topic of study. The Natural Language Toolkit (NLTK) is a Python module for NLP [3]. The first step is counting and tokenizing the content in the document. The program data flow follows the following steps during the algorithm execution time.

The first step is to input the selected data and involves the section of the features that one wants to check on particular cases, such as the number of paragraphs in the document. The whole idea of the process will be established via the Naïve Bayes Classifier. True or False probabilities and percentages are determined by classifying the new data. Fake news if the likelihood score is greater than zero, or actual news otherwise. The number of verbs in the article, the number of entities, and the number of double quotes mentioned, will be used to compute the score. Below is the algorithm used in the implementation process of this study.

Fake News Detection

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

In [2]: path = "./data/"
        true_df = pd.read_csv(path + 'True.csv')
        fake_df = pd.read_csv(path + 'Fake.csv')

In [3]: true_df['label'] = 0

In [4]: fake_df['label'] = 1

In [5]: true_df.head()
```

Figure 1:

Loading datasets

Data Flow Process

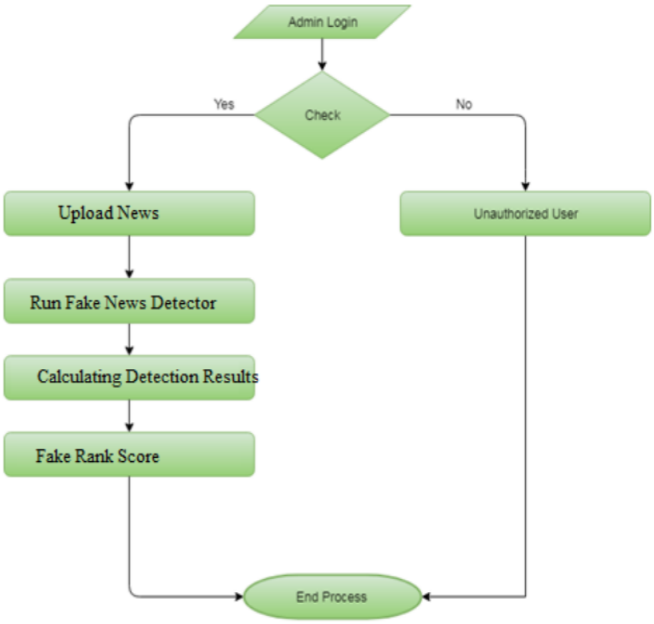

Figure 2

Data flow process

The algorithm will then look for quotation marks, such as single or double quotes, in the text or paragraph of the paper. Once the stop words and standard terms have been removed, the Naive Bayes algorithm will take effect. It uses the Naive Bayes classifier to categorize the document's text and calculates the probability and score. The content in the paper will be labeled as either false or authentic news based on its score in the scoring system.

The proposed algorithm performs several tests on the data supplied during the trial step before the deployment. The algorithm prompts the python IDE to bring the NumPy library into the current environment in use. The as np section of the code then instructs Python to assign NumPy the alias of np. This enables users to utilize NumPy functions by just entering np. The few tests performed during the deployment of the proposed algorithm show that some of the articles scored returned the value "0" and others "1," indicating that the news is genuine and false, respectively. The table depicts the results of the tested program;

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

In [2]: path = "./data/"
        true_df = pd.read_csv(path + 'True.csv')
        fake_df = pd.read_csv(path + 'Fake.csv')

In [3]: true_df['label'] = 0

In [4]: fake_df['label'] = 1

In [5]: true_df.head()
```

Out[5]:

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0

```
In [6]: fake_df.head()
```

Out[6]:

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

Figure 3

Labeling

Results of the Tested Data

Out[6]:

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

```
In [7]: true_df = true_df[['text','label']]
        fake_df = fake_df[['text','label']]

In [8]: dataset = pd.concat([true_df , fake_df])

In [9]: dataset.shape

Out[9]: (44898, 2)
```

Figure 4

Results after concatenation

Classifier Performance Metrics

Multinomial NB

```
In [35]: from sklearn.naive_bayes import MultinomialNB

In [36]: from sklearn.metrics import accuracy_score,classification_report

In [37]: clf = MultinomialNB()

In [38]: clf.fit(train_data, train_y)
         predictions  = clf.predict(test_data)

In [39]: print(classification_report(test_y , predictions))
```

	precision	recall	f1-score	support
0	0.95	0.95	0.95	3247
1	0.96	0.95	0.96	3753
accuracy			0.95	7000
macro avg	0.95	0.95	0.95	7000
weighted avg	0.95	0.95	0.95	7000

Now predict on both train set

```
In [40]: predictions_train = clf.predict(train_data)
         print(classification_report(train_y , predictions_train))
```

	precision	recall	f1-score	support
0	0.96	0.95	0.96	13458
1	0.96	0.96	0.96	14542
accuracy			0.96	28000
macro avg	0.96	0.96	0.96	28000
weighted avg	0.96	0.96	0.96	28000

Figure 5

Performance metrics of Multinomial NB classifier

PROPOSED ALGORITHM

In the proposed algorithm, it makes use of the custom attribution feature to do the classification. An attribution is specified using the conceptual principles; it employs the following mathematical expression. Let C be a random resource span of length len(C) for attributing. The attribution span is the distance in character spaces between the beginning and the conclusion of the content span, represented by the "d" character. As a result, the mathematical formula is as follows;

Equation 1: Cue identification

$$(Source, Cue) \leq x^i + len(C) + 2d$$

$$\exists (Source, Cue) for\ C\ s.t. \{ \quad or \quad (Source, Cue) \geq x^i \qquad (1)$$

The forward and tail attribution spans are searchable sub-spans of the attribution. Once this tool has been created, it may search the forward trail attribution space and categorize the quotation. The custom classifier feature used named entity recognition algorithms to look for the named entities or people or organizations that may be ascribed as having produced a quotation to identify the source.

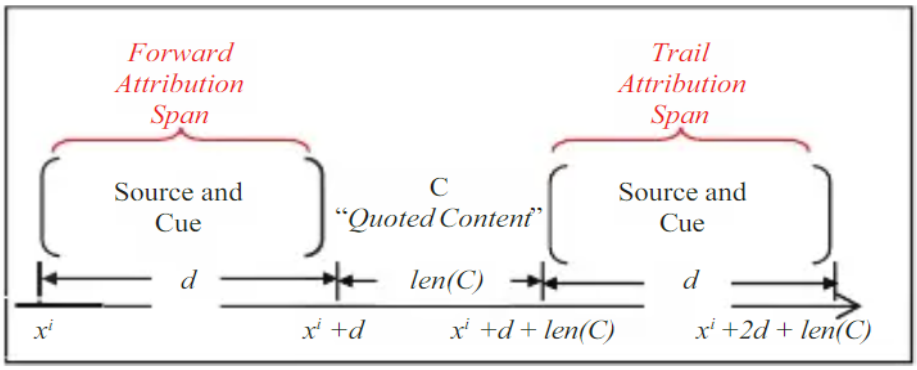

Figure 6

Attribution span and absolute distance

The approach of deploying NLP to detect fake news would be a wise idea but would incredibly face challenges involving synonyms. Synonyms, like contextual knowledge, may pose problems since we employ a variety of words to communicate the same concept. Some of these terms may convey the same meaning. In contrast, others may be degrees of complexity (little, little, tiny, minute), and various individuals employ synonyms to express somewhat different meanings within their lexicon [2]. Consequently, while developing NLP systems, it is critical to incorporate all potential definitions and synonyms for a term. Although text analysis algorithms may still sometimes make errors, the more relevant training data they acquire, the better they understand synonyms. Also, text analysis may be hampered by incorrectly spelled or misconstrued terms. While grammar checkers and autocorrect can handle most errors, they don't always know what the author means. A computer may have difficulty understanding spoken language because of mispronunciation, varied accents, and a stuttering lexicon [2]. However, these difficulties may be reduced as language databases expand and intelligent assistants are educated by their unique users.

Table 1

Classification report of test data with predictions

	precision	recall	f1-score	support
0	0.95	0.95	0.95	3354
1	0.96	0.95	0.95	3646
accuracy			0.95	7000
macro avg	0.95	0.95	0.95	7000
weighted avg	0.95	0.95	0.95	7000

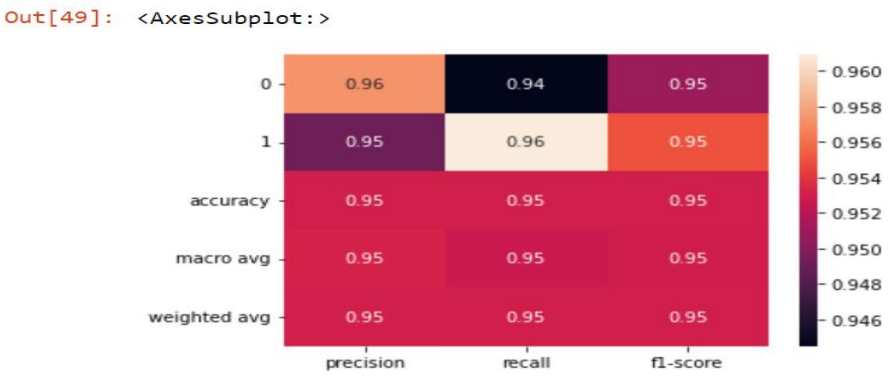

Figure 7

Heat map of classification report of test data

Table 2

Classification report of train data with predictions

	precision	recall	f1-score	support
0	0.96	0.96	0.96	13300
1	0.96	0.95	0.96	14700
accuracy			0.96	28000
macro avg	0.96	0.96	0.96	28000
weighted avg	0.96	0.96	0.96	28000

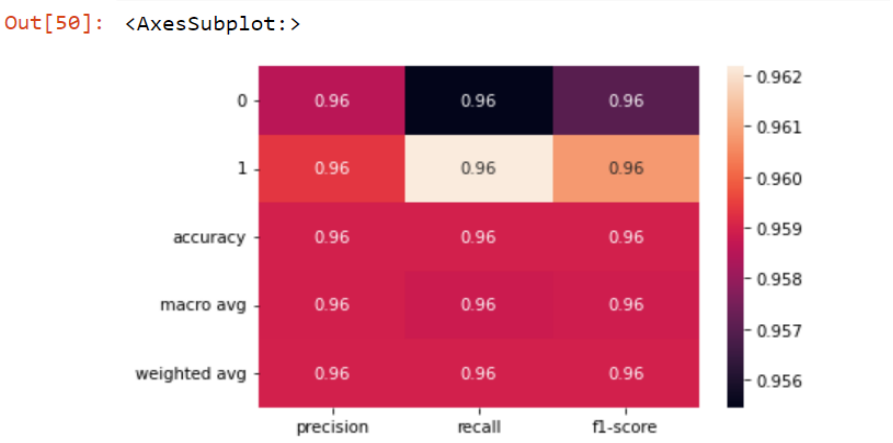

Figure 8

Heat map of classification report of train data

CONCLUSION

An algorithm for detecting false news has been developed, and the findings of that work are detailed in this paper. Fake news has become one of the most pressing issues we face today. As a result, we find it more difficult to tell if the information we get is genuine or fraudulent. A fake news detection system was created to combat the spread of false information. Real or fake news may be classified by this approach, which also provides a proportion of the news found to be false. Natural language processing (NLP) and machine learning methods such as Multinomial Naive Baye's classifier were employed to develop this system. First, a model is trained using the proper training datasets, and then it is tested using the correct testing datasets. The algorithm will determine whether or not a piece of information is genuine or not based on the likelihood and score it receives. NLP would be an effective strategy to embrace to fight against false information being supplied on the internet. As a result, the credibility of the information, trust, and information reputation will be controlled, and help know trusted sources.

BIBLIOGRAPHY

[1] G. Tolios, "How I created a fake news detector with python," *Medium*, Oct. 08, 2021. https://towardsdatascience.com/how-i-created-a-fake-news-detector-with-python-65b1234123c4

[2] M. Arumugam, "Processing the textual information using open natural language processing (NLP)," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3361108

[3] E. J. Rifano, Abd. C. Fauzan, A. Makhi, E. Nadya, Z. Nasikin, and F. N. Putra, "Text summarization menggunakan library natural language toolkit (NLTK) berbasis pemrograman python," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 2, no. 1, pp. 8–17, Apr. 2020, doi: 10.28926/ilkomnika.v2i1.32

[4] C. A. Watson, "Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development," *International Journal of Legal Information*, vol. 46, no. 2, pp. 93–96, Jul. 2018, doi: 10.1017/jli.2018.25

[5] T. Traylor, J. Straub, Gurmeet, and N. Snell, "Classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator," *IEEE Xplore*, Jan. 01, 2019. https://ieeexplore.ieee.org/document/8665593 (accessed Mar. 26, 2020).

[6] M. G. Yadav, R. Nennuri, N. Sairam, Y. S. Teja, and G. Prasad, "Classifying fake news articles using natural language processing and supervised learning estimator," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 6847–6856, Jun. 2021, Accessed: Jan. 08, 2022. [Online]. Available: https://www.annalsofrscb.ro/index.php/journal/article/view/6781/5099