Detection of Textual Propaganda Using Passive Aggressive Classifiers

P. Niranjan Kumar[1]
[1]Assistant Professor, Keshav Memorial Institute of Technology,
Hyderabad, India, niranjan.1216@gmail.com

ABSTRACT

Nowadays, social media activity, particularly news that spreads over the network, is a major source of knowledge. People search out and chew up news from internet-based living because of the low effort, easy access, and rapid dissemination of information. Twitter, as one of the most well-known continuing news sources, also happens to be one of the most dominant news disseminating media.

It has already been known to wreak significant harm by disseminating snippets of gossip. Online clients are typically susceptible, and everything they do on web-based networking media is assumed to be trustworthy. As a result, automating counterfeit propaganda detection is critical to maintaining a vibrant online media and informal organization. In order to computerize propaganda news identification in Twitter datasets, this research develops a technique for recognizing propaganda text messages from tweets by figuring out how to anticipate precision evaluations. This paper proposes a supervised machine learning technique, Passive aggressive classifiers that uses Count Vectorizer and Term Frequency-Inverse Document Frequency Vectorizer as feature extraction to detect propaganda news based on the polarity of the corresponding article. Finally, this algorithm uses dataset with 43000 records and shows good accuracy.

Key words: Count Vectorizer, Term Frequency - Inverse Document Frequency, Passive Aggressive Classifier.

1. INTRODUCTION

Social networks have increased access to valuable information, and their decentralized and unregulated environment has allowed misinformation to spread widely. The huge propagation of misinformation on social media has turned into a global hazard, subtly impacting public opinion and endangering social and political progress. In recent years, misinformation detection has been a popular study area. It is observed that new research issues and approaches in detecting misinformation have received a lot of attention as a valuable and continuously emerging research subject.

Propaganda is an oxymoron that smears the credibility of reporting that achieves the verifiability criteria and serves the public good. Additionally, with social networks, misinformation spreads faster, deeper, and wider. Because of the free flow of information through social networking sites like Twitter, controlling the dissemination of false or misleading news is one of the most difficult tasks in this electronic era. The Suspicious content is typically available in text and is used in the majority of cases to carry out suspicious activities.

As a result, one of the most complicated tasks ahead of us is to create a system that can efficiently identify suspicious text from specified resources. Americans believe that 65 percent of the news they receive on social media is propaganda, according to the latest study by the Knight Foundation. Researchers from (Zhou et al., 2019) discovered that propaganda is becoming more prevalent over time. Thus, it is necessary to identify phony news. Furthermore, people have a limited ability to separate non-propagandist text from the propagandist text given a large number of information we are exposed to when utilizing social media.

Misinformation is a huge social concern because it undermines the integrity of legitimate information, endangering democracy, justice, the economy, public health, and security. Moreover, we want to build a technique can that be applied for classification problems to measure the insight about the COVID-19 pandemic and identify some common misunderstanding among the wider public, which will assist us in informing public health agencies, and then better strategies for educating the public and making them understand not to fall for these myth scan be devised.

The spreading of misinformation not only forces people to accept false views and alters how they respond to the truth but also jeopardizes the entire information ecosystem's credibility. The use of social media by adversaries to distribute false or misleading information constitutes a political hazard. For instance, according to a survey, approximately one-third of residents in the United States, Spain, Germany, the United Kingdom, Argentina, and South Korea have seen erroneous or misleading information about COVID-19 on social media.

Fake news has a significant impact on democracy. The American presidential election demonstrated how it

destabilizes and distorts people's opinions. Since fake news is created to deceive readers, detecting it just based on the content of the news is challenging. Because news information varies in terms of style and subject, it becomes necessary to implement an effective method for detecting lies and propaganda. Because misinformation can have a heavy impact in a matter of minutes, detecting them during an early stage is vital. Researchers have recently become very interested in detecting propaganda.

2. RELATED WORK

The paper "Detection of Bangla Fake News using MNB and SVM Classifier" states Fake news detection from English texts and other languages has gotten a lot of attention, but just a few studies have been done in Bangla. Because this topic still requires a lot of attention, this work offers an experimental investigation on the detection of Bangla false news on social media. They employed Count Vectorizer and Term Frequency - Inverse Document Frequency Vectorizer as feature extraction. In this study to detect Bangla fake news using two supervised machine learning techniques, Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) classifiers. The method used in this paper detects bogus news based on polarity of article [4].

In this article "Detecting Fake News using Machine Learning and Deep Learning Algorithms" author tells In the context of computerizing forged news detection in Twitter datasets, a model for spotting faked news messages from twitter postings by figuring out how to anticipate precision evaluations. After that, they compared five well-known Machine Learning techniques, including Support Vector Machine, Nave Bayes Method, Logistic Regression, and Recurrent Neural Network models, to show how efficient the classification performance on the dataset was. The SVM and Nave Bayes classifiers outperformed the other algorithms in their experiments.[7].

The article "Identifying propaganda from online social networks during COVID19 using machine learning techniques" illustrates that during COVID-19, many propagandistic messages concerning the fatal virus are being distributed. They used Twitter's application programme interface (API) to extract data, then they manually annotated it. The most relevant features are chosen using hybrid feature engineering. Machine learning algorithms are being used to categorize tweets into binary categories. Among all the algorithms, the decision tree produces the best results. Feature engineering might be refined for better outcomes, and deep learning could be employed for classification tasks.[5]

The evaluation article of "Analytics of machine learning-based algorithms for text classification" shows The performance of several machine learning algorithms on different datasets is studied and compared[3].According to the paper "A Survey on Computational Propaganda Detection," this new form of self-expression has had unintended consequences, the most evident of which is that society is now open to future threats emanating from a number of sources. The topic drew a lot of attention in 2016, a year marked by

unprecedented levels of micro targeted online disinformation and misinformation, particularly in the aftermath of Brexit and the 2016 presidential election. In 2020, the COVID-19 pandemic spawned the first worldwide infodemic, which happened to coincide with the US Presidential election. People tended to lower their expectations since the information was dependable, thanks to the dissemination of bogus material posing as news. They had a better understanding of knowledge when it came from a range of sources.[8].

The Paper "Detection of propaganda using logistic regression" tells Various propaganda techniques, such as the use of logical fallacies or appealing to the emotions of the audience, are used to manipulate people's viewpoints in order to promote a preconceived goal. In this study author suggests Logistic Regression-based technique that automatically classifies whether a statement is propagandistic or not. To distinguish these two groups, the features such as the TF-IDF, BERT vector, sentence length, readability grade level, emotion feature, LIWC feature, and emphatic content feature. The combination of linguistic and semantic variables yields an F1 score of 66.16 percent, which is significantly higher than the baseline.[6]

Propaganda campaigns attempt to change people's minds in order to further a specific agenda. They take advantage of the Internet's confidentiality, social networks' micro-profiling capabilities, and the ease of organizing and maintaining synchronized networks of records to reach billions of dollars of social network users with persuasive communication that are directly aimed to topics that each individual user is sensitive to, ultimately influencing the outcome on a targeted issue. The paper "Detecting fake news using machine learning: A systematic literature review" presents a review of state of the art in computational propaganda detection from the perspectives of Natural Language Processing and Network Analysis in this survey, recommending that these fields should work together.[1]

In both academia and the industry, building automated disinformation detection tools has become a hot topic. WhatsApp is one of the most common forms of misinformation in several developing nations, including Brazil, India, and Mexico. Despite this, due to WhatsApp's encrypted messaging nature, there is few disinformation detection algorithms created expressly for this network. The paper "Fake WhatsApp. br: NLP and machine learning techniques for misinformation detection in brazilian portuguese WhatsApp messages" uses FakeWhatsApp.BR application, which contains a collection of WhatsApp conversations in Brazilian Portuguese that was manually categorized and obtained from Brazilian public groups. In addition, a tested a collection of 108 misinformation detectors that combined Natural Language Processing-based feature extraction methods with a variety of well-known machine learning algorithms.[2]

A considerable number of troll accounts have appeared in recent years in an attempt to sway public opinion on social media sites. They are frequently active in disseminating false

info, fake news, and propaganda in order to distract and foment conflict. This study focuses on detecting troll tweets in both English and Russian, assuming that they are generated by a "troll farm. We reduce these risks to assessing whether a solitary tweet was written by a "troll farm" by reducing it to an authenticity authentication problem "whether you have an account or not We use different machine learning methods, including deep learning, to evaluate a supervised classification approach with monolingual, cross-lingual, and bilingual training scenarios.[9].

2.1 Limitations in Existing Work:

There are variety of classification machine learning techniques for detecting propaganda, including decision tree, Support Vector Machine(MNB), Multinomial Naive Bayes(MNB).However this models have some drawbacks. The cost of training of SVM model, for example, is limited. When training huge datasets, SVM takes lengthy time. It necessitates hyper-parameter tuning, which is not simple and time consuming.

MNB works well with tiny datasets, however this may lead to lower accuracy. MNB also only works with snippet text, assuming that each characteristic is independent of the others. MNB assumes that features are unrelated to one another, but this may not always be the case. Choosing a kernel function is complicated and time consuming in case of SVM, and also for

MNB probabilities aren't precise, and it doesn't allow for feature interaction.

.

3. PROPOSED METHOD

3.1 Passive Aggressive Classifier:

Passive-Aggressive algorithms are generally used for large-scale learning and used to perform classification tasks. It is one of the few online-learning algorithms. Passive-aggressive algorithm is an incremental learning algorithm. Passive-aggressive algorithm is particularly effective and efficient for certain applications. It operates by acting passive for accurate classifications and aggressive for any inaccuracies in incremental machine learning algorithms as shown in Figure1. This algorithm trained incrementally by datasets either alone or consecutively or in small groups termed mini-batches and the model is gradually upgraded.

This algorithm used to build a model which is trained and deployed in production in online learning in such a way that it continues to learn as new data sets come. As a result, we can safely assume that an algorithm like the Passive-Aggressive Classifier is best suited for systems that receive data in a continuous stream.

This is highly beneficial in circumstances where there is a large amount of data and training the full dataset is computationally impossible due to the sheer bulk of the data. An online learning algorithm will simply obtain a training data set, update the classifier, and then discard the data set. Instead of inspecting all of the training samples at once, incremental learning inspects one at a time. The obvious benefit is that you have a small memory footprint, which is a significant benefit.

A very good example of a Passive-aggressive algorithm would be to detect propaganda on asocial media website like Twitter, where new information is added every second. The amount of data needed to dynamically read data from Twitter on a continual basis would be enormous, thus an online-learning algorithm as shown in Figure 4, would be great.

Passive: Keep the model and don't make any changes if the prediction is right. i.e., the data in the case is insufficient to generate any model adjustments.

Aggressive: Make modifications to the model if the prediction is wrong. i.e., a change to the model could correct the problem.
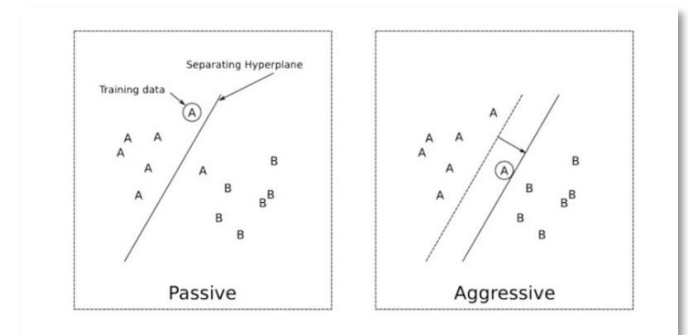
Figure 1: Passive Aggressive Classifier

The temporal dimension has been designated by the index t. The samples can, in fact, continue to arrive indefinitely in this instance. Of course, if they are selected from the same data-generating distribution, the algorithm will continue to learn (likely without significant parameter changes), but if they are obtained from a completely different distribution, the weights will gradually forget the prior one and learn the new one. This algorithm works with a binary categorization based on bipolar labels.

$$X = \{\overline{x_0}, \overline{x_1}, \overline{x_2} \dots \dots, \overline{x_t}, \dots.\} \; where \; \overline{x_i} \in R^n$$

$$Y = \{y_0, y_1, \dots \dots, y_{t,\dots\dots}\} where \; y_i \in \{-1, +1\}$$

The value of L is bounded between 0 (meaning perfect match) and K depending on f(x(t),)with K¿0 (completely wrong prediction).

A Passive-Aggressive algorithm works generically with this update rule:

$$\widetilde{y}_t = sign(\overline{w}^T.\overline{x}_t)$$

$$L(\overline{\theta}) = \max(0, 1 - y.f(\overline{x}_t; \overline{\theta})$$

$$\overline{w}_{t+1} = argmin_{\overline{w}} \; \frac{1}{2} \, \|\overline{w} - \overline{w}_t\|^2 + C\xi^2$$

$$L(\overline{w}; \overline{x}_t, y_t) \le \xi$$

To understand this rule, let's assume the slack variable =0 (and L constrained to be 0). If a sample x(t) is presented, the classifier uses the current weight vector to determine the sign. If the sign is correct, the loss function is 0 and the argmin is w(t). This means that the algorithm is passive when a correct classification occurs.
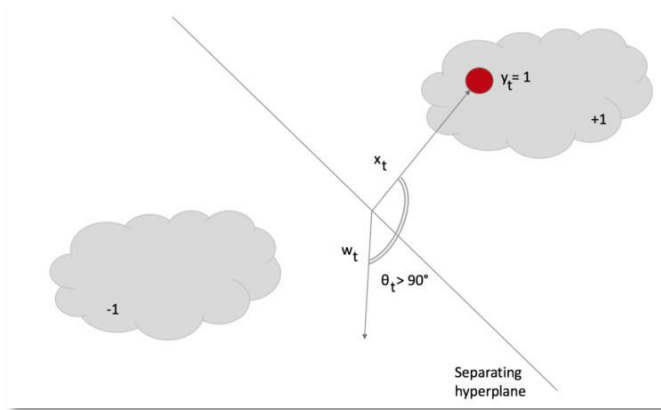
Let's now assume that a misclassification occurred:

Figure 2: Hyper plane Separation

The angle ¿ 90°, therefore, the dot product is negative and the sample is classified as -1, however, its label is +1 as shown in Figure 2. In this case, the update rule becomes very aggressive, because it looks for a new w which must be as close as possible as the previous (otherwise the existing knowledge is immediately lost), but it must satisfy L=0 (in other words, the classification must be correct).

The introduction of the slack variable allows to have soft-margins (like in SVM) and a degree of tolerance controlled by the parameter C. In particular, the loss function has, allowing a larger error. Higher C values yield stronger aggressiveness (with a consequent higher risk of destabilization in presence of noise), while lower values allow a better adaptation. In fact, this kind of algorithms, when working online, must cope with the presence of noisy samples (with wrong labels). A good robustness is necessary; otherwise, too rapid changes produce consequent higher misclassification rates.

After solving both update conditions, we get the closed-form update rule:

This rule confirms our expectations: the weight vector is updated with a factor whose sign is determined by y(t) and whose magnitude is proportional to the error. Note that if there's no misclassification the nominator becomes 0, so w(t+1) = w(t), while, in case of misclassification, w will rotate towards x(t).In the next figure, the effect has been marked to show the rotation, however, it's normally as smallest as possible.

This rule confirms expectations: the weight vector is updated with a factor whose sign is determined by y(t) and whose magnitude is proportional to the error. Note that if there's no misclassification the nominator becomes 0, so w(t+1) = w(t), while, in case of misclassification, w will rotate towards

x(t).In the Figure 3, the effect has been marked to show the rotation, however, it's normally as smallest as possible.
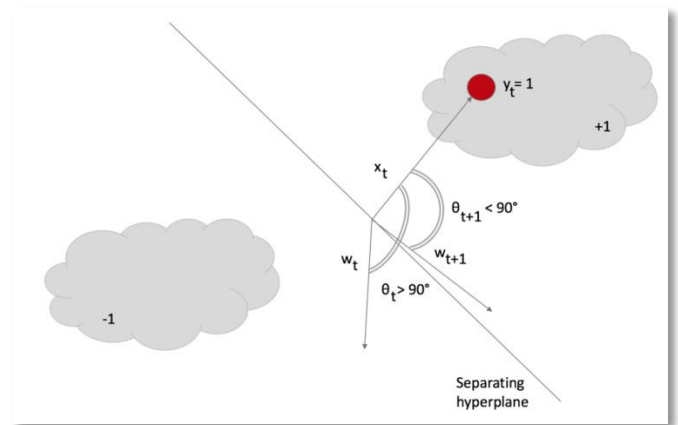
Figure 3: Hyper plane
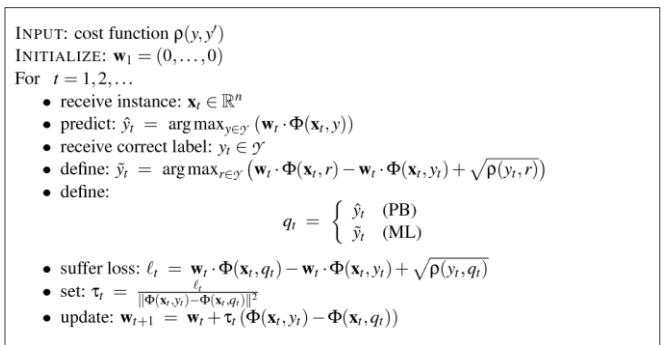
Figure 4: Algorithm

3.2 System Architecture

The architecture in Figure **5** describes various steps to build a system to detect propaganda message using passive aggressive classifiers algorithms. The following steps are followed to build a model.

1. Data Collection
2. Data Preprocessing.
3. Feature Selection and Extraction.
4. Build model using passive aggressive classifiers.
5. Prediction.
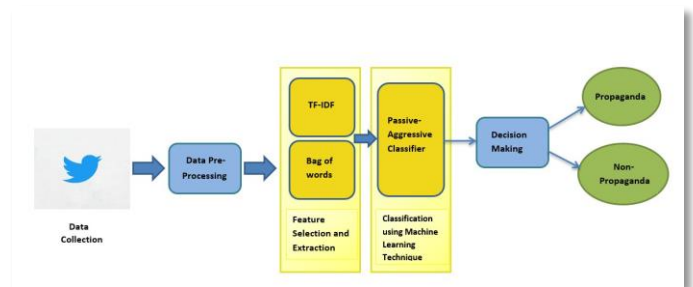
Figure 5: Architecture of Passive Aggressive Classifier

3.4 Dataset:

The dataset consists of 43000 records. The CSV file contains the News Title, News Text, Subject and Date as shown in Figure 6, 7 and Figure 8 describes about data flow diagram

Figure 6: Non-Propagandist Dataset
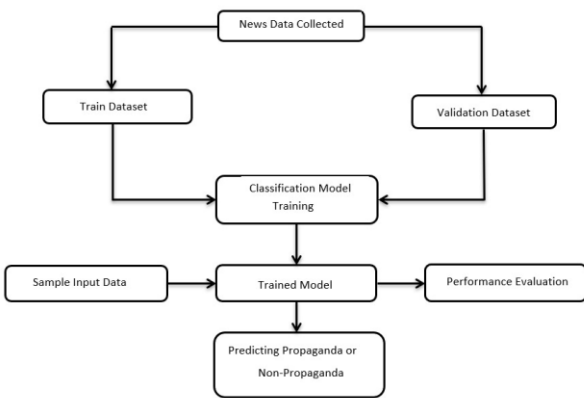

Figure 7: Propagandist dataset

Figure 8: Data Flow Diagram

4. IMPLEMENTATION

The following steps are followed to build model using python

Step 1: pd.read is used to read CSV file and the resulted output will be in form of dictionary.

filename = pd.read csv('/content/Filename.csv')

Step 2: This function is used to determine a Machine Learning Accurateness.

sklearn.metrics.accuracy_score (predict labels, actual labels)

Step 3: This method is used to plot the confusion matrics with the given parameters

plot_confusion_matrix(cm, classes,normalize=False,title='Confusion matrix',cmap=plt.cm.Blues):

Step 4: This method is used to divide the dataset into two parts: training and testing.

X_train, X_test, y_train, y_test = train_test split(X, y, test size=0.2, random state=0)

Step 5: The functions .lower and .split are used to change text to lower case and split them into tokens.

text = text.lower(), text = text.split()

Step 6: This function is used to execute the application file

app.run()

Step 7: The above code is used to create a text field that will take the user's input text and predict it.

original_text = TextAreaField('Original Text', validators=[Length(min=20, max=10000)], render kw='placeholder': 'Enter text here..')
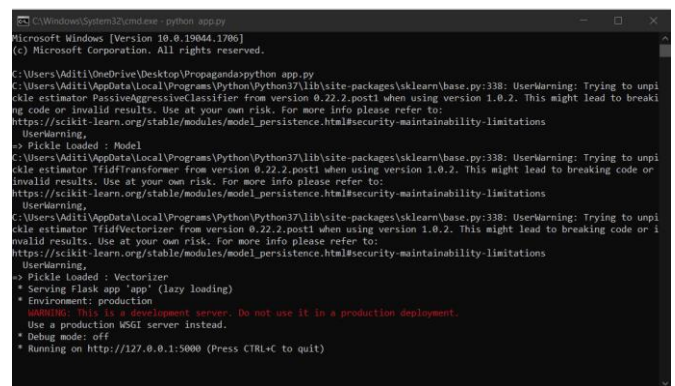

Figure 9: Console Screen

Step 8: To run project we need to execute the command " python app.py" in prompt and URL is generated in Console screen shown in Figure 9, open the URL in browser the Figure 10 shows how the interface looks like.

Type any random text for the classification. We can also generate random text by clicking on"Generate Text" button as shown in Figure 11.
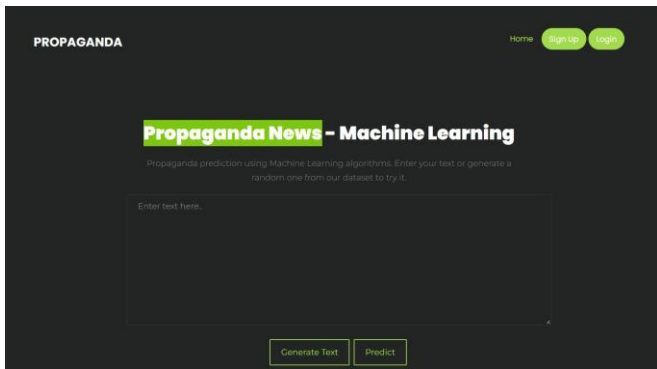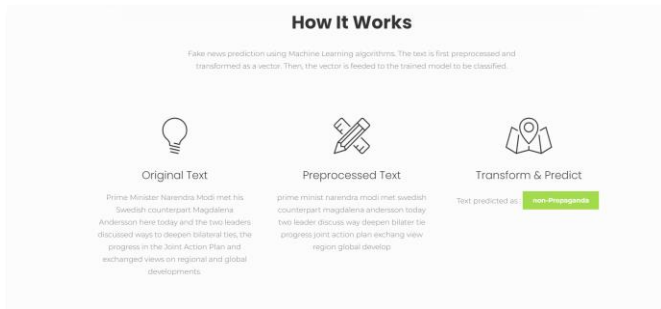

Figure 10: Main Screen

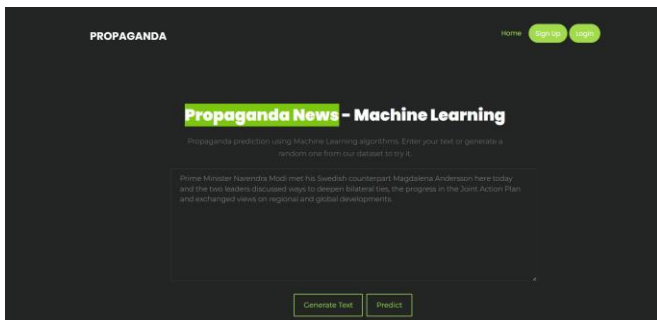**Figure 11:** How It Works

Figure 12: Output Screen-01

Then text input is given to the model as shown in the Figure 12.

Figure 13: Output Screen - 02

In the above screen (Figure 13) On clicking on "**Predict**" it shows us as "**NON-PROPAGANDA**".

4.1Evaluation of Model:

The accuracy of Passive-Aggressive Classifier is: 95.47

Confusion Matrix: Figure 14 shows confusion matrix drawn on test data.

Figure 14: Confusion Matrix

From the above analysis we can state that confusion matrix gives high accuracy than any other machine learning algorithm.

5. CONCLUSION

Due to the obvious widespread use of the internet, it is now simpler to propagate propaganda. On most of the platforms, users have no restrictions when it comes to posting news on these platforms. As a result, few people make use of these platforms to distribute propaganda about individuals or organizations. This research concentrated on the task of detecting propaganda and created an intelligent method based on several useful features. It uses features include Tokenization, Lemmatization, term frequency/inverse document frequency (TF-IDF), and Bag of Words (BoW). A novel classification technique proposed to improve the identification of propaganda on social networking websites, with this research come to the conclusion that on our dataset, the passive-aggressive classifier performs considerably better than Multinomial Nave Bayes. The combination of these features with the passive-aggressive classifier achieved the best performance.

The accuracy score is 95.47%. The passive-aggressive classifier is ideal for large-scale data sets. When the number of features for each data set is more than the number of training data sets, the passive-aggressive classifier does not get affected by it and will perform significantly well. The 'zero-frequency problem' occurs when the Multinomial Naive Bayes algorithm gives zero probability to a categorical variable whose type in the test data set was not accessible in

the training dataset, in the passive-aggressive classifier model, this was easily overcome. We had data sets of nearly 43,000 news articles to work within our research. To increase the performance, we plan to implant the features we designed in this model and study new features from propaganda approaches. We also intend to continue adding more datasets in the future and also to increase the number of features and lexicons of this research.

REFERENCES

1. Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. **Detecting fake news using machine learning: A systematic literature review**. arXiv preprint arXiv: 2102.04458, 2021.

2. Lucas Cabral, Jos´e Maria Monteiro, Jos´e Wellington Franco da Silva, C´esar Lincoln C Mattos, and Pedro Jorge Chaves Mourao. **Fakewhastapp. br: NLP and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages**.In Proceedings of the 23rd International Conference on Enterprise Information Systems,ICEIS, pages 26–28, 2021.

3. Sayar Ul Hassan, Jameel Ahamed, and Khaleel Ahmad. **Analytics of machine learning based algorithms for text classification. Sustainable Operations and Computers,** 3:238–248, 2022.

4. Md Gulzar Hussain, Md Rashidul Hasan, Mahmuda Rahman, Joy Protim, and Sakib Al Hasan. **Detection of bangla fake news using mnb and svm classifier**. arXiv preprint arXiv:2005.14627, 2020.

5. Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. **Identifying propaganda from online social networks during covid-19 using machine learning techniques.** International Journal of Information Technology, 13(1):115–122, 2021.

6. Jinfen Li, Zhihao Ye, and Lu Xiao. **Detection of propaganda using logistic regression.** In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 119–124, 2019.

7. Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, et al. **Detecting fake news using machine learning and deep learning algorithms.** In 2019 7th International Conference on Smart Computing and Communications (ICSCC), pages 1–5. IEEE, 2019.

8. Giovanni Da San Martino, Stefano Cresci, Alberto Barr´on-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. **A survey on computational propaganda detection.** arXiv preprint arXiv:2007.08024, 2020.

9. Lin Miao, Mark Last, and Marina Litvak. **Detecting troll tweets in a bilingual corpus**. In Proceedings of the 12th Language Resources and Evaluation Conference, pages