International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

WELFAKE – WORD EMBEDDING OVER LINGUISTIC FEATURES FOR FAKE NEWS DETECTION

Barath M[1], Sangeethkumar C[2], Naveen N[3], Karthickram S[4], Mr. Partha Sarathi P[5]
[1] *UG – Electrical and Electronics Engineering, Bannari Amman Institute of Technology, Sathy, Erode.*
[2] *UG – Electrical and Electronics Engineering, Bannari Amman Institute of Technology, Sathy, Erode.*
[3] *UG – Electrical and Electronics Engineering, Bannari Amman Institute of Technology, Sathy, Erode.*
[4] *UG – Electrical and Electronics Engineering, Bannari Amman Institute of Technology, Sathy, Erode.*
[5]*GUIDE – Computer Science Engineering, Bannari Amman Institute of Technology, Sathy, Erode.*

ABSTRACT
News is the only mode and set of information that helps the public to know what's happening everyday globally. We have started our path of reading news digitally, by which many "Fake news" are being circulated. Fake news is false or misleading information presented as news. Fake news often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. People unknowingly believe those fake news as original one without any analysis or study. Since the machine cannot read the words we use, we are going to use “ML model” to train our dataset to the machine. Our project is a two-phase benchmark model named WELFake based on word embedding where each and all words are converted into numerical values which is further processed to classify based on certain matching property using machine learning. The first phase preprocesses the data set and validates the veracity of news content by using linguistic features. The second phase merges the linguistic feature sets with WE(Word Embedding) and applies voting classification. The classification is based on words and meaning matching and this matching percentage should be above a threshold value we fix. In this paper we are going to discuss about choosing the best algorithm based on our needs and accuracy and complete the task successfully.
Keywords: Word Embedding, classification, vectorizing, matrix formation, common features.

INTRODUCTION
News is the only mode and set of information that helps the public to know what's happening everyday globally. We have started our path of reading news digitally, by which many "Fake news" are being circulated. Fake news is false or misleading information presented as news. Fake news often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. People unknowingly believe those fake news as original one without any analysis or study. Since the machine cannot read the words we use, we are going to use “ML model” to train our dataset to the machine.

In certain venues, fake news these days is causing a variety of problems, from sarcastic articles to manufactured news and deliberate government propaganda. In our society, fake news and a lack of faith in the media are serious issues that have far-reaching effects. The concept of "fake news" has recently changed due to the blathering social media discourse. Obviously, a narrative that is intentionally misleading is fake news. Some of them now use the phrase to discount the evidence that conflicts with their favoured worldviews.

The phrase "fake news" became widely used to refer to the problem, especially when describing pieces that were published primarily in order to generate revenue from page views but contained factual errors and misinformation.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

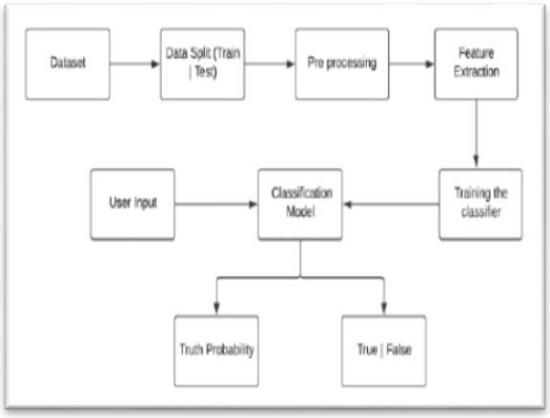
EXISTING SYSTEM

Machine learning techniques for fraud detection have been the subject of extensive research, with the majority of it concentrating on categorising online reviews and publicly accessible social media posts. The issue of identifying "fake news" has received a lot of attention in the literature, especially since late 2016 during the American Presidential election.

They point out that superficial parts-of-speech (POS) tagging and straightforward content-related n-grams have frequently failed to account for crucial context information, rendering them ineffective for the classification task. These techniques have only been shown effective when used in conjunction with more sophisticated analytical techniques.

PROPOSED SYSTEM

In this paper, a model is built based on the vectorizer. Implementing a best accuracy classifier that will be the best and standard for text-based processing since this challenge is a type of text classification. The real objective is deciding which type of text to utilise and constructing a model for text transformation (count vectorizer vs tfidf vectorizer) (headlines vs full text).

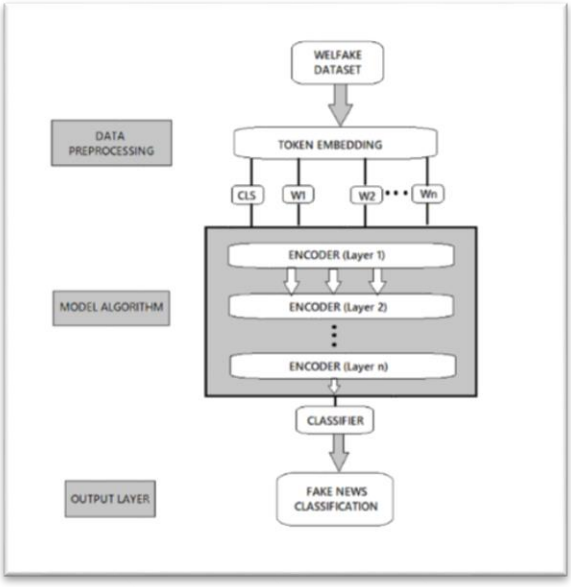


So, there must be two parts to the data-acquisition process, “fake news” and “real news”. Collecting the fake news was easy as Kaggle released a fake news dataset consisting of 13,000 articles published during the 2016 election cycle. Now the later part is very difficult. That is to get the real news for the fake news dataset. It requires huge work around many Sites because it was the only way to do web scraping thousands of articles from numerous websites. With the help of web scraping a total of 5279 articles, real news dataset was generated, mostly from media organizations.

REQUIREMENTS:

- Python
- numpy
- pandas
- PortStemmer
- Sklearn

FLOW DIAGRAM:



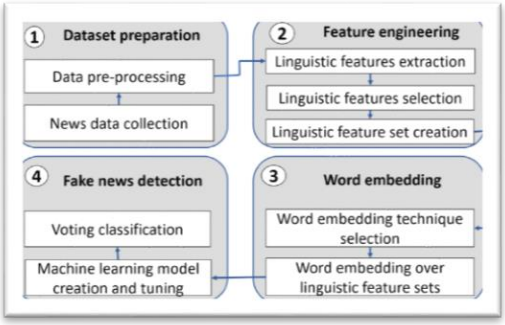
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

WELFAKE MODEL:

WELFAKE generally means to embed words that are given to the system to numerical coordinates and use them to make certain comparisons based on our algorithm and needs.

There are four steps to be done in the cumulative model:

- Generates a bigger data set with better generalization.
- Recognizes the twenty most important language traits and generates three distinct LFS depending on categories;
- Two WE approaches are used to train various ML models; and
- Uses a two-stage vote classification to get the final forecast.



A. DATA SET PREPARATION:

1) News Data Collection:

This is critical for a fair and impartial data collection, as well as for giving high-quality training data and producing good outcomes. Although there are several accessible data sets available for the investigation of false news, the research has revealed major limitations in terms of quantity, category, or bias. We produced a more complete WELFake data set after extensive research that integrates four data sets, Kaggle, McIntire, Reuters, and BuzzFeed, for two reasons. For starters, they share a framework with two categories (i.e., real and fake news). Second, by integrating the data sets, the limits and biases of each individual data collection are reduced.

- The number of short sentences (less than 10 words) indicating actual news outnumbers those expressing false news.
- The text readability of false news is lower than that of true news.
- Fake news pieces are more subjective than actual news items.
- The number of articles representing true news outnumbers those representing fake news.

2) Data Processing:

This uses many approaches, depending on the data collection and aims, to tackle various difficulties in the acquired data, such as typographic mistakes, unstructured data format, and other constraints.

- Missing data addresses the presence of undefined (NaN) and blank values (NULL) in the data collection, which impedes the feature engineering process. Because removing missing value data items may result in the loss of critical information, we used a missing value imputation technique that guesses missing values and then examines the whole data set as if these values were the real observed ones.
- Inconsistent data differs from other data points due to errors made during data gathering. For outlier detection and correction, we used a variety of visualization approaches and mathematical functions, including the box plot, scatter plot, Z-score, and inter-quartile range (IQR) score.
- Duplicate data, or deduplication, reduces redundancy, which can lead to biased conclusions when the data is gathered by the same individual.
- Stop words (and other noise) that make the phrase grammatically full but have no semantic relevance in news categorization processes are removed by removing irrelevant data. Removing stop words while maintaining important tokens only improves model performance considerably.
- Stemming transforms text into its root word by using the Porter-Stemmer algorithm on text characteristics to increase accuracy. If the root word is not recognized, it constructs the canonical form of a corresponding word.

B. FEATURE ENGINEERING:

1) Linguistic Features Extraction: This is the process of converting raw text into data for the ML algorithm. Feature extraction seeks to provide a feature set that summarizes the information in the

DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

original data collection, hence speeding up model training and improving data visualization and learning accuracy. WELFake collected 87 text-based linguistic characteristics from cutting-edge publications, categorizing them as syntactic (i.e., writing pattern, quantity) or semantic (i.e., grammar, psycho-linguistics).

- The writing pattern highlights the text's writing style by utilising sentence kinds, determinants, special characters, and modifiers.
- Grammar is concerned with the text readability index, as well as word structure, average syllables per word, easy word use ratio in a word list, and sentence complexity.
- Based on semantic and subjective, psycho-linguistic evaluates text sentiment and information opinion.
- Quantity detects speech information pieces by counting the number of verbs, adjectives, adverbs, syllables, and words, as well as the frequency with which adjectives, adverbs, and words appear in a phrase.

2) Linguistic Features Selection:

This is the process of selecting key characteristics for data classification, which reduces the number of input features, lowers computing costs, and improves the ML model's accuracy. We evaluated the Pearson's correlation coefficient of each feature with the other features in the same category for this purpose and excluded those with a correlation coefficient more than 0.7, indicating a strong positive linear connection [23] between the two features. According to Occam's razor and the minimal description length notion [62], a learning model with fewer characteristics is simpler and more exact. We repeated this approach until all of the minor characteristics were removed. Table V summarizes the remaining 20 most important WELFake properties, which are organized into four groups.

3) LFS Creation:

This divides the 20 linguistic characteristics into various sets, allowing several WE approaches for unbiased model training. To get a clear result from subsequent vote categorization, we require an odd number of input sets. As a result, we generated at least three unique LFS based on four categories.

- The sentence structural complexity of any text is defined by the readability index. Based on the readability index, we determine the level or grade of the text writer, which aids in assessing the news as true or false. As a result, we dispersed the three readability index features evenly across the three LFS and assigned one feature to each group.
- As discussed, psycho-linguistic characteristics play a crucial role in detecting bogus news. As a result, we used all three in all three LFS
- Because quantity characteristics play a role in news categorization, we allocated them evenly among the three LFS.
- Because the writing pattern has five characteristics, we evenly allocated three features to each LFS.

C. WORD EMBEDDING:

1) WE Selection:

Because ML techniques cannot directly handle plain text, this finds the best suited way for transforming plain text into a numeric value. In the literature, we discovered two popular WE categories: 1) content-based, such as term frequency-inverse document frequency (TF-IDF) and count vectorizer (CV), which are based on prior information, and 2) context-based, such as Word2Vec, GloVe, and FastText, which are based on written text patterns. We chose the content-based WE that focuses on writing patterns rather than context since false news writers (fakesters) tend to repeat similar terms.

- **CV:** Also known as one-hot encoding, CV turns a text document into a histogram vector, with each element representing the number of times the word appears in the document. The length of the vector is determined by the number of unique words in the corpus.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

- **TF-IDF:** This is a more advanced form of CV that displays the significance of a term (representing a word) in a corpus with its presence in the document. The term frequency tf(t,d) that computes the occurrence of a term t in a document d and the inverse document frequency idf(t,D) that computes the relevance of that phrase t in a corpus of documents D are multiplied.

2) WE over LFS:

This enhances output prediction because specified characteristics may not always forecast well and require extra training approaches. We integrated the WE and LFS for this purpose. On the three LFS, we used the TF-IDF and CV WE methods and discovered that CV produces superior results. We attained a maximum accuracy of 95.61% using SVM on CV, and a maximum accuracy of 95.12% using bagging. As a result, in Section VII-B, we chose CV and paired it with LFSs for further accuracy examination of several models.

D. FAKE NEWS DETECTION:

1) ML Model Creation and Tuning:

The LFS with WE is processed using six ML methods: SVM, NB, KNN, DT, Bagging, and AdaBoost. We tested each ML model on random samples of the WELFake data set with four training-testing data combinations: 60%-40%, 70%-30%, 80%-20%, and 90%-10%. To increase accuracy, we manually tuned the six separate models using the hyperparameters shown in Table VI. We iterated through several hyperparameter value combinations from the specified feasible value ranges until we achieved a state-of-the-art accuracy of at least 96%.

2) Voting Classification:

This method use ensemble learning to collect prediction outputs from different models and delivers a result that minimises error and over-fitting. In general, there are two techniques to voting classifiers: soft voting based on probability and hard voting based on maximum votes. Because false news identification is a binary classification problem, we employ hard voting, which predicts a target variable Y based on the maximum votes mode provided by several models Mi to a class:

$Y=\text{mode}\{M1(X),M2(X),\ldots,Mn(X)\}$

E. WELFAKE – FAKE NEWS DETECTION ALGORITHM:

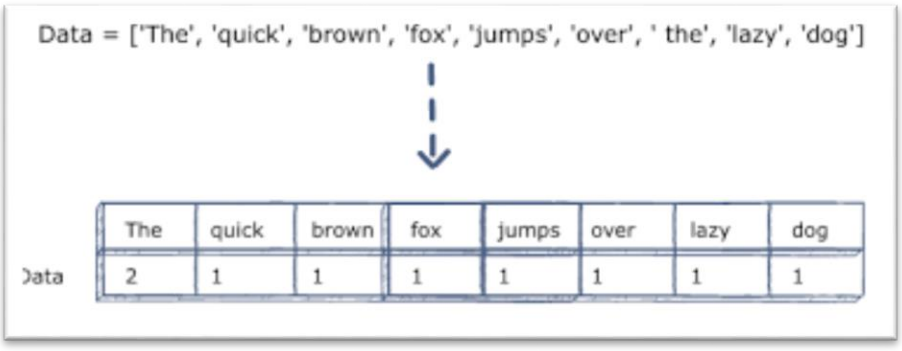


Fig. A simple demonstration of how vectorizer works

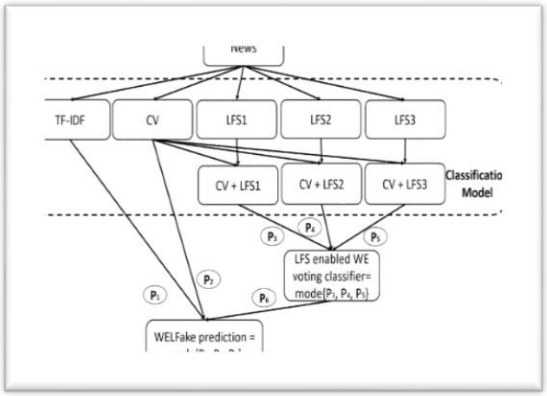
WELFake is a binary news classification approach (as real or fake) that employs TF-IDF and CV, three linguistic feature sets (LFS1, LFS2, and LFS3), and a hard voting classifier. The WELFake workflow is divided into four steps:

- Apply TF-IDF and CV to the complete data set, record the results in P1 and P2, and choose the best WE based on accuracy.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

- CV (better performing approach) is applied to three well-defined LFS and the results are saved in P3, P4, and P5.

- Apply the WE hard voting classifier to P3, P4, and P5, and output the prediction P6.

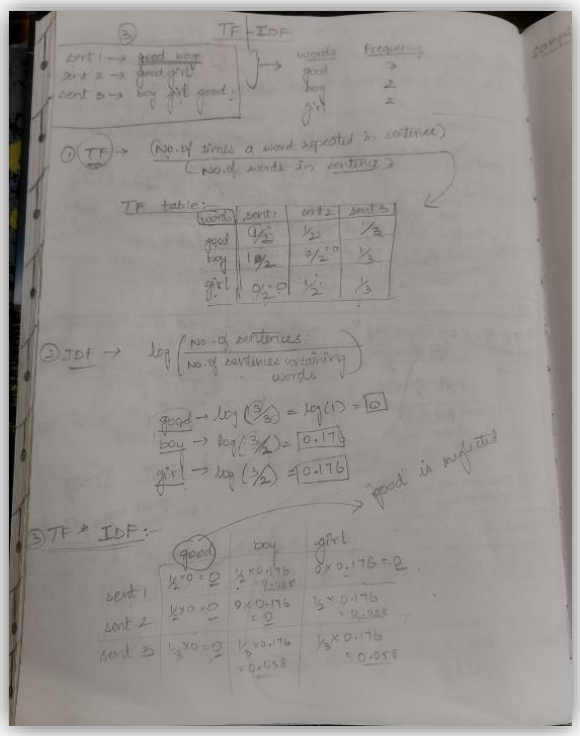


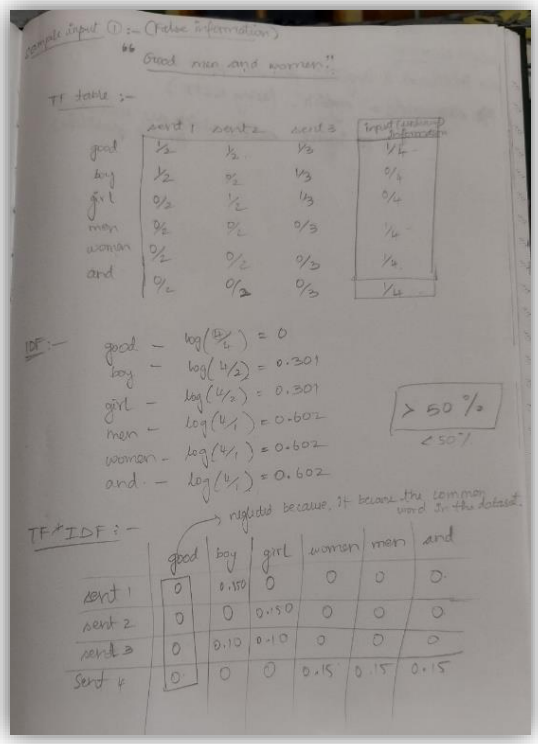
- P1, P2, and P6 are combined using the hard voting classifier to give the final prediction result.

The Algorithm works by the following methodology:

1. WELFake data gathering in line 1 is followed by data set preparation in line 2.
2. Line 3 extracts linguistic characteristics from the data set, while line 4 applies Pearson's coefficient to identify the relevant linguistic features. Line 5 generates an odd number of LFS for voting.
3. WE: (i.e., CV and TF-IDF) (i.e., CV and TF-IDF) Lines 6 and 7 apply to the complete data collection. Line 8 is where we choose the optimal technique to combine with the multiple LFS established in line 5. Lines 9 and 10 use the LFS in conjunction with the best WE technique.
4. It trains the data sets from line 5 on several ML classification models and selects the best results from each set. Line 12 provides the hard voting output by applying the voting classifier to the results obtained on the various LFS using the best ML classification model. Line 13 runs the hard voting classifier, CV and TF-IDF (line 7) again, and produces the final news categorization prediction.

Fig. Calculation on how vectorizer algorithm works inside the system:





International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

CLASSIFIER:

1) Random Forest

A group of decision trees is referred to as a "Random Forest" under trademark law. We have a collection of decision trees—hence the name "Forest"—in Random Forest. Each tree assigns a classification to a new object based on its qualities, and we say the tree "votes" for that classification. The classification with the highest votes is selected by the forest (over all the trees in the forest). A classification system made up of several decision trees is called the random forest. It attempts to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree by using bagging and feature randomness when generating each individual tree. Like its name suggests, a random forest is made up of numerous independent decision trees that work together as an ensemble. The class with the highest votes becomes the prediction made by our model. The random forest's individual trees each spit forth a class prediction. The random forest model performs better than any of its component models when it operates as a committee of many generally uncorrelated models (trees), which is why it does. How then
does random forest make sure that each tree's behaviour is not overly connected with any of the other trees in the model?
It uses two methods:

1.1 Bagging (Bootstrap Aggregation)

Considering how sensitive decision trees are to the data they are trained on, even minor adjustments to the training set can result in radically different tree architectures. By enabling each individual tree to randomly sample from the dataset with replacement and produce various trees as a consequence, random forest takes advantage of this. This method is often referred to as bootstrapping or bagging.

1.2 Feature Randomness

When splitting a node in a typical decision tree, we analyse all potential features and choose the one that results in the greatest gap between the observations in the left node and those in the right node. In contrast, only a random subset of features are available to each tree in a random forest. This drives even more variety among the model's trees, which ultimately leads to decreased correlation between them and increased diversification.

2. Logistic Regression

It is a classification algorithm rather than a regression one. Based on a set of independent variables, it is used to estimate discrete values (binary values like 0/1, yes/no, and true/false) (s). In plain English, it determines the likelihood that an event will occur by fitting data to a logit function. It is also known as logit regression as a result. Given that it forecasts probability, its output values range from 0 to 1. (as expected). In mathematics, the predictor variables are combined linearly to represent the log probability of the outcome.

Odds = p/(1-p) = probability of event occurrence / probability of not event occurrence ln(odds) = ln(p/(1-p)) logit(p)=ln(p/(1-p))= b0+b1X1+b2X2+b3X3....+bkXk

3.Passive Aggressive Classifier

The online algorithm known as the passive aggressive algorithm is excellent for categorising huge data streams (e.g. twitter). It moves quickly and is simple to apply. By using an example, learning from it, and then discarding it, it functions [24]. In the event of an incorrect classification, such an algorithm remains passive but becomes aggressive, updating and adjusting. It does not converge, unlike the majority of other algorithms. Its goal is to make updates that fix the loss while barely changing the weight vector's norm.

CONFUSION MATRIX:

A table known as a confusion matrix is frequently used to describe how a classification model, also known as a "classifier," performed on a set of test data for which the true values were known. It enables the visualisation of an algorithm's performance. A classification problem's prediction outcomes are compiled in a confusion matrix. Count values are used to describe the number of accurate and inaccurate predictions for each class. This is the confusion matrix's secret. The confusion

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

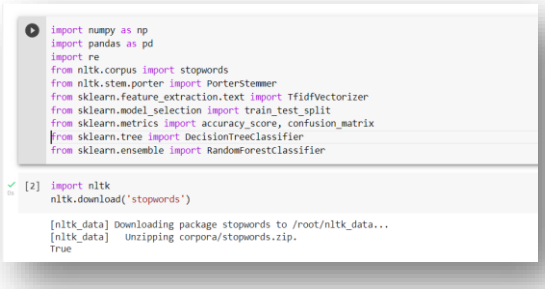
matrix demonstrates the manner in which your classification model makes predictions while being confused. It provides insight into a classifier's errors, but more crucially, the types of errors that are being made.

Total	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	TP	FN
Class 2 (Actual)	FP	TN

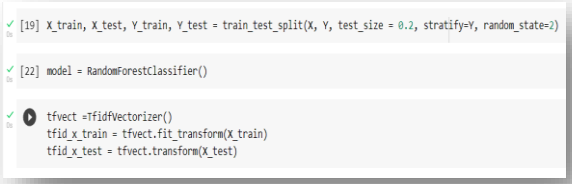
CODE IMPLEMENTATION:

- Libraries to import was Numpy, Pandas, Stopwords, PorterStemmer, Vectorizer, Train_test_split, Accuracy_score, Confusion_matrix, Classifier.
- Import dataset using pandas
- Use PorterStemmer to clean the dataset
- Use train_test_split method to split the data as X_train, X_test, Y_train, Y_test
- Update the data set to classifier. Use classifier to predict the output.
- Create the function to accept the given input
- Pass the input to function to predict whether the news as real or fake.

These are the libraries which are used in our project



- PorterStemmer is used for data cleaning in our given dataset.
- Vectorizer is used for converting a Text to Numerical Co-ordinates.
- Classifier is used for predictio



Train_Test_split is used for splitting a data and update a data into x_train, Y_train, X_test, Y_test. TfidVectorizer is used as a vectorizer.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

Stemming process will remove the unwanted data such as is, was, So, stopwords will contain set of word and these words are removed from our dataset. It will allow only alphabetical letter other than that all are removed.

RESULT AND DISCUSSION:

Total=7819	Fake (predicted)	True (predicted)
Fake (Actual)	538	95
True (Actual)	124	510

Random Forest Classifier

Total=7819	Fake (predicted)	True (predicted)
Fake (Actual)	475	158
True (Actual)	140	494

Passive Aggressive Classifier

The majority of tasks are completed online in the twenty-first century. Applications like Facebook, Twitter, and online news stories are replacing newspapers, which were formerly favored as tangible copies. The forwards on WhatsApp are another important source. Fake news is a growing issue that only complicates matters and seeks to sway people's opinions and attitudes. In favor of using digital technologies. We created a system called Fake News Detection that classifies user input as either real or false. Various NLP and machine learning techniques must be employed to do this. A suitable dataset is used to train the model, and its performance is also assessed using a variety of performance metrics. The model we created using Random Forest gives accuracy of 82%. Logistic Regression also gives accuracy around 80%. Passive Aggressive Classifier gives efficiency around 76%. Without using data cleaning Passive Aggressive Classifier gives higher efficiency but using data cleaning process Random Forest and Logistic Regression gives good efficiency.
But to use this as a Real time product Passive Aggressive Classifier is best suit. It is an online learning algorithm classifier. It can train the model regularly with updated news. No need to train the entire dataset.

CONCLUSION:
We have to stop believing all the false news we hear and be secure and aware on what's happening around us. Hence using these kinds of technologies would help us to detect the false news spreading like a virus. Our project would surely help us in finding out those false news and destroy them from public.

REFERENCES:
1. W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, "Trust evaluation in online social networks using generalized network flow," IEEE Trans. Comput., vol. 65, no. 3, pp. 952–963, Mar. 2016.
2. M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami,M. M. Hassan, and G. Fortino, "Credibility in online social networks: A survey," IEEE Access, vol. 7, pp. 2828–2855, 2019.
3. S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, "Facilitating time critical information seeking in social media," IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2197–2209, Oct. 2017.
4. [Z. Zhang, R. Sun, X. Wang, and C. Zhao, "A situational analytic method for user behavior pattern in multimedia social networks," IEEE Trans. Big Data, vol. 5, no. 4, pp. 520–528, Dec. 2019.
5. M. Schudson and B. Zelizer, "Fake news in context," in Understanding and Addressing the Disinformation Ecosystem. Philadelphia, PA, USA: Annenberg School for Communication, Apr. 2017, pp. 1–4.

DOI:10.46647/ijetms.2022.v06i06.080 ISSN: 2581-4621

6. S. Zaryan, "Truth and trust: How audiences are making sense of fake news," M.S. thesis, Media Commun. Studies, Lund Univ. Publications Student Papers, Stockholm, Sweden, Jun. 2017. [Online]. Available: https://lup.lub.lu.se/student-papers/search/publication/8906886
7. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.
8. R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, "A large-scale study of the Twitter follower network to characterize the spread of prescription drug abuse tweets," IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1232–1244, Dec. 2019.