Fake Information Detection

Suraj Joshi
Student
Dept of MCA
Dr.Ambedkar Institute of Technology
Bangalore, India
joshisuraj9999@gmail.com

Staline Edson Dsouza
Student
Dept of MCA
Dr.Ambedkar Institute of Technology *Bangalore, India*
staline2012@gmail.com

Dr. Bharathi S
Professor
Dept of MCA
Dr.Ambedkar Institute of Technology
Bangalore, India
Bharathishivu_s@yahoo.co.in

Shivaleela S
Assistant Professor
Dept of MCA
Dr.Ambedkar Institute of Technology
Bangalore, India
shivaleela.joteppa@gmail.com

***Abstract*— Technology has been the evolution in human history, but simultaneously, underlying fact is that there is no distinction between trustable and non-trustable sites. Now, anyone can publicize contents, reliable or unreliable, that can be compelling to the networked users globally. Unfortunately, misinformation pile-ups to an exceptional interest on the network, mostly on social-networking sites.**

I. Introduction

As time flows, the amount of data, especially text data increases exponentially. Along with the data, our understanding of AI also increases and the computing power enables us to train very complex and large models faster. Fake information has been gathering a lot of attention worldwide recently. The effects can be political, economic, organizational, or even personal. This paper discusses the approach of natural language processing and machine learning in order to solve this problem. Use of bag-of-words, n-grams, count vectorizer has been made, TF-IDF, and trained the data on five classifiers to investigate which of them works well for this specific dataset of labelled information statements. The precision, recall and f1 scores help us determine which model works best.

Along with the increase in the use of social media platforms like Facebook, Twitter, etc. information spread rapidly among millions of users within a very short span of time. The spread of fake information has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates.

The goal of this project was to create a tool for detecting the language patterns that characterize fake and real information through the use of machine learning and natural language processing techniques. The results of this project demonstrate the ability for machine learning to be useful in this task.

A. What is Fake Information

Fake Information is information designed to deliberately spread hoaxes, propaganda and disinformation. It is a form of exploitative journalism; Fake Information encases information that may possibly be confidence tricks in addition to and in generally be diffused through social sites as well as other networked mass media. This is frequently done to foist certain concepts and it is normally attained with activists' daily activities. These kinds of information may contain phoney i.e., untruthful and/or exaggerated claims, and may end up going as viral information beside algorithms, and in addition public may land in a new filter-bubble.

B. Why is Fake Information Detection so Important

Contemporary world has become acceptable and the entire mankind have to be grateful for the enormous benefaction of internet and technology for communication, transmission and data sharing.

Technology has been the evolution in human history, but simultaneously, underlying fact is that there is no distinction between trustable and non-trustable sites. Now, anyone can publicize contents, reliable or unreliable, that can be compelling to the networked users globally. Unfortunately, misinformation pile-ups to an exceptional interest on the network, mostly on social-networking sites.

C. Objective

The three most prevalent motivations for writing fake information and chosen only one as the target for this project as a means to narrow the search in a meaningful way. The first motivation for writing fake information, which dates back to the 19th century one-sided party newspapers, is to influence public opinion. The second, which requires more recent advances in technology, is the use of fake headlines as clickbait to raise money. The third motivation for writing fake information, which is equally prominent yet arguably less dangerous, is satirical writing.

Machines are better at detecting and keeping track of statistics than humans, for example it is easier for a machine to detect that the majority of verbs used are "suggests" and "implies" versus, "states" and "proves." Additionally, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources. Either of these methods could prove useful in detecting fake information, but we decided to focus on how a machine can solve the fake information problem using supervised learning that extracts features of the language and content only within the source in question, The current project involves utilizing machine learning and natural language processing techniques to create a model that can expose documents that are, with high probability, fake information articles. Many of the current automated approaches to this problem are centred around a “blacklist” of authors and sources that are known producers of fake information. But, what about when the author is unknown or when fake information is published through a generally reliable source? In these cases, it is necessary to rely simply on the content of the information article to make a decision on whether or not it is fake. By collecting examples of both real and fake information and training a model, it should be possible to classify fake information articles with a certain degree of accuracy. The goal of this project is to find the effectiveness and limitations of language-based techniques for detection of fake information through the use of machine learning algorithms including but not limited to convolutional neural networks and recurrent neural networks. The outcome of this project should determine how much can be achieved in this task by analysing patterns contained in the text and blind to outside information about the world. This type of solution is not intended to be an end-to end solution for fake information classification. Like the “blacklist” approaches mentioned, there are cases in which it fails and some for which it succeeds. Instead of being an end-to-end solution, this project is intended to be one tool that could be used to aid humans who are trying to classify fake information. Alternatively, it could be one tool used in future applications that intelligently combine multiple tools to create an end-to-end solution to automating the process of fake information classification.

D. Existing-System

Existence of significant researches typically concerning machine-learning providing possible solutions for fraudulent information detection, nearly all being centred on sorting through on-line presents and overviews., and widely accessible social-media marketing blogposts. In view of overdue in year-2016 through US Presidential-election, regard to the issue to figuring out phoney information. Conroy, Rubien, and Chien [6.] presents an overview with methods that is encouraging to attain accuracy to deceiving articles, remembering that the content-associated n-grams and parts-of-speech (POS) labelling have been demonstrated lack of data categorization, tend to fail with information framework. Contrarily, all these methods have evidently shown effectiveness when designed with more analysis involved.

E. Proposed System

In the presented project, a prototype is designed using count-vectorizer or a tf-idf matrix (i.e., word counts related to the defined articles in the dataset). Given a fact that the issue is textual set, the naive Bayes classifier has been implemented to process the texts as this algorithm is best suited for text-based processing. The primary objective is to design a paradigm to transform texts (count-vectorizer vs tf-idf vectorizer) and deciding the type/kind of text to be used (headlines or full-text). Forthwith, the further step is to obtain the most optimum properties for count-vectorizer and/or tf-idf-vectorizer, this can be done by considering the ngram-number of the most-used terms, and/or phraseologies, casing or not, most importantly eliminating the stop-words which are commonly used words such as “as”, “in”, and “to” and by using the words that occur at least a certain number of times in the textual-Dataset.

F. Machine Learning

Natural Language processing or Machine-learning is an area in applied computer-science that allows computers to publish programs on its own without having to be programmed clearly. The key goal of machine learning techniques is to build algorithm that obtains input and makes use of some analysis to predict the outcome within a range. That mainly makes choices by detecting styles from the earlier data and generalizing it in the future data. Machine learning is mainly classified into Supervised Learning, Unsupervised Learning and Reinforcement Learning.

1) SUPERVISED LEARNING

It’s a type of machine-learning technique where we are going to have insight variable x and output variable y which algorithm produce an event that will predict the results based on the input values. Supervised learning is mainly classified into regression and classification. The supervised learning problems is presumed to be regression problems, where, if the result value is a continual value such as height, weight and so on. Similarly, a supervised problem is considered a category problem if the output value is a distinct ideal such as man or female, disease or disease free and so on. Few of the classification algorithms involve support vector machines (SVM), neural network systems, decision trees, naive bayes classifier, K nearest neighbours, so on. Some of the regression algorithms include nonlinear regression, linear regression, neural network systems and decision trees.

2) UNSUPERVISED LEARNING

This is a type of machine-learning algorithm wherever; we extract effects through the input info without containing branded responses. Unsupervised understanding algorithms are labelled directly into cluster research and association. A new cluster analysis is usually a one wherever object which can be related to the other person usually are grouped together. Relationship rule is typically the one where typically the existence of exciting relationships involving the parameters in the dataset is discovered. A number of the examples of unsupervised learning algorithms usually are hierarchical clustering, K-means clustering, gaussian blend models, self-organizing routes, hidden Markov designs, Apriori algorithm in addition to so on.

3) SEMI SUPERVISED LEARNING

Semi-Supervised learning technique in machine-learning, lies somewhere in between supervised and unsupervised learning, as this approach uses both labelled and unlabelled data- input for training- generally, a bit of labelled-data and aggregate of unlabelled-data. The system applies this-one method to substantially upgrade learning precision. Generally, semi- supervised learning is adopted when the obtained labelled data requisite skilled and related resources so as to train or learn from it. If not, obtaining unlabelled data mostly does not require supplementary resources.

4) REINFORCEMENT LEARNING

This type of algorithm which allows equipment to learn its behaviour from the comments that is obtained from the atmosphere. A few of the algorithms that come under Encouragement learning are Q-learning, Temporal difference, Heavy Adversarial Networks and so on.

II. LITERATURE SURVEY

Recent political events have led to an increase in the popularity and spread of fake information. As demonstrated by the widespread effects of the large onset of fake information, humans are inconsistent if not outright poor detectors of fake information. With this, efforts have been made to automate the process of fake information detection

A. *The Influence And Effect Of Fake Information*

The information highway or the world wide web is mainly motivated by commercial advertising [1]. Sites with interesting and full-frontal (sensational) statements exist incredibly popular, in addition contributes to advertising agencies make capital out of the high-rise visitors to the website [1]. It was eventually uncovered that the initiator of inaccurate information and websites could generate income by way of computerized advertising that recompense traffic to the websites. Typically, the question continues to exist that how inaccurate information would then affect people. Typically, the increase of false information may cause dilemma and uncalled stress amongst public [2]. Inaccurate information that is intentionally developed to deceive which might also damage the public represents digital disinformation [3]. Disinformation gets the potential to cause issues, within minutes, for lots of people [2]. Deceptions has been well known to interrupt process of elections, create distress, debates and antagonism among the public [3].

B. *Fake Information On Social Media*

Nowadays, the social networking possesses an important part of our day to day lives [6]. Traditional ways of acquiring information possess almost vanished in order to cover the method for social media press platforms [6]. In 2017, it had been documented that Facebook was the greatest social networking system, providing more than 19 million customers globally [6]. A part of Facebook was spreading the Fake Information possibly offering the biggest effect from all of the other social media platforms [13]. It had been outlined that 44% of global customers receive their information from Facebook [13]. Facebook customers of around 23%, have specified that they have discussed mis-information, possibly knowingly or not really. The spread of Fake Information will be powered by interpersonal networking platforms in fact it is the reality.[13].

III. Methodology

For more simple and common NLP classification tasks, such as sentiment analysis, there is an abundance of labeled data from a variety of sources including Twitter, Amazon Reviews, and IMDb Reviews. Unfortunately, the same is not true for finding labeled articles of fake and real news. This presents a challenge to researchers and data scientists who want to explore the topic by implementing supervised machine learning techniques. Implementation is done in five sub stages:

1. Assortment of information from platforms like Reddit and Twitter
2. Selection of unique structures for arrangement and training project
3. Assessment of various model’s efficiency depending on extracted structures
4. Performance Improvements

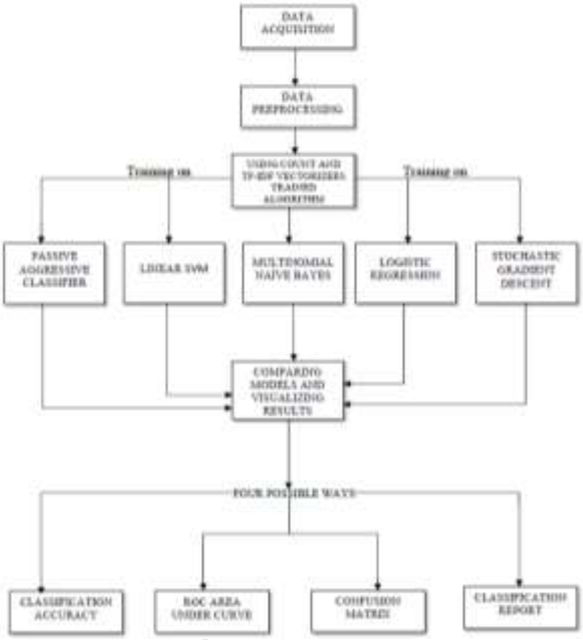


Fig 1. Flow Chart of the System

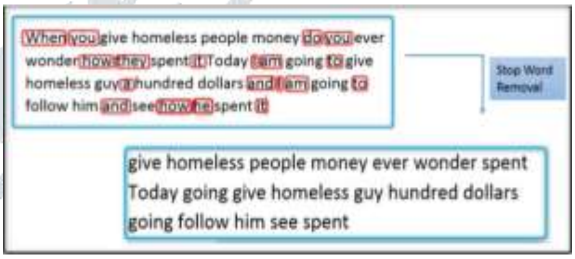
IV. Implementation

A. *TEXT PREPARATION*

Data will always have some sort of impurities in it may it be stop words or punctuations or tokens. To get the maximum efficiency the data is cleaned of all such impurities before the model gets this data as input, this is also called as preprocessing the training data.

This step was comprised of Alteration to single case: First step in preprocessing is to convert all the words into one single case eighter uppercase or lowercase such that repeating of the same word in a different case is avoided

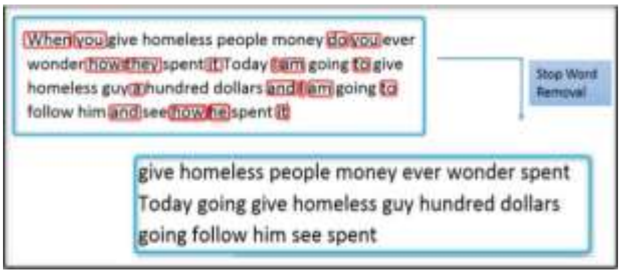
1. Punctuations Removal: Punctuation marks do not make a significance while processing inscribed text data. So, removal of punctuations help reduce the overall size of text.


When you give homeless people money do you ever wonder how they spent it Today I am going to give homeless guy a hundred dollars and I am going to follow him and see how he spent it

give homeless people money ever wonder spent Today going give homeless guy hundred dollars going follow him see spent

Fig 2. Punctuation Removal

2. Stop words Removal: These are the words that will be frequently used. Some of the words include, a, the, of, at, e.g. ,on etc. They provide a definite design for the text structure but not to context. In case they are to be treated as unique structure, poor performance will come as outcome. Therefore, Stop words are removed from training data

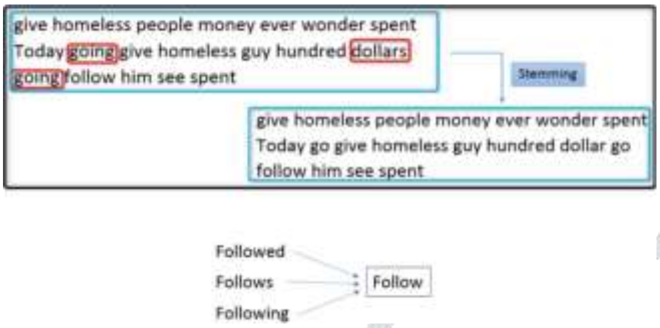

When you give homeless people money do you ever wonder how they spent it Today I am going to give homeless guy a hundred dollars and I am going to follow him and see how he spent it

give homeless people money ever wonder spent Today going give homeless guy hundred dollars going follow him see spent

Fig 3. Stop-word Removal

3. Tokenization: This is to separating the written characters into a arrangement of terms / number of terms / group of words. It is done to get vectors based on frequency values that are acquired because of the tokens.

4. Stemming: This is a method used for removing prefix and suffix from word, ending with the stem. Inflectional and derivational forms can be reduced by using this technique which forms the base for the word.



give homeless people money ever wonder spent
Today going give homeless guy hundred dollars
going follow him see spent

give homeless people money ever wonder spent
Today go give homeless guy hundred dollar go
follow him see spent

Fig 4.Illustrating Stemming

5. Lemmatization: This changes words into the word's root. Using vocabulary, morphological analysis is done to pick root word. Lemmatization was executed to enhance values of vectors based on frequency.

B. REPRESENTATION OF WORD-VECTOR

For this objective, Word Embedding methods are employed for converting textual content to vectors, in order for the system to process all data.

1. Bag of Word: It is a technique considers every article as documents and computes the frequency for every single word in the specific document. This can be used to create numerical statistic data representation. It is also known as fixed length vector-features.
2. Word-Embedding: This format usually tries to chart / map a word to the vector utilizing a dictionary. The next rate of recurrence-based word embedding vectors was utilized for training the data.
3. Count Vector as a unique structure: This is the matrix interpolation associated with datasets, in which the row of data symbolizes the documents in the corpus, column of data symbolizes a feature in the corpus, cell of data represents count number of a specific feature in the document
4. TFIDF vectors as a feature: TFIDF weight provides the representation of relative significance of terms in a document and in throughout corpses.
   a) Term Frequency (TF): It is the count of the frequency of a term appearing within a document. A high value defines that a term has appeared frequently than others, and hence, the text is a fine match when the it comes to search using words or specific terms.

$$TF(t,\ d) = \frac{Number\ of\ times\ t\ occurs\ in\ a\ document\ 'd'}{total\ word\ count\ of\ document\ 'd'}$$

   b) Inverse Document Frequency (IDF): It is the words that occurs as many times in a document as in other documents, which might also be irrelevant. E.g., etc., "a", "an", "the", "on" IDF measures how substantial a term is within the entire corpus.

$$IDF(t) = log_e \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ in\ it}$$

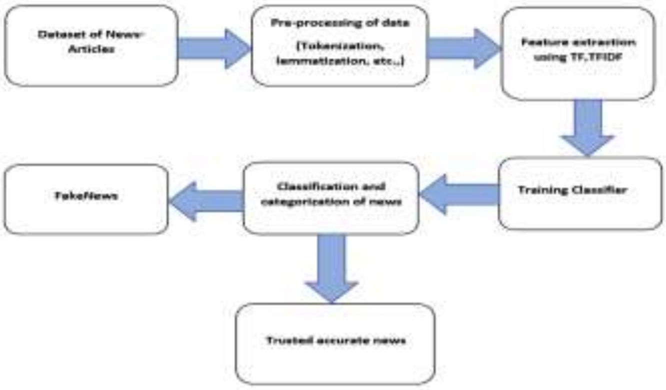
   c) Term Frequency Inverse Document Frequency (TFIDF): TFIDF works through chastising the word with the highest frequency and assigning less weights and giving higher weights to words, which are present as a proper subset to the corpus, and has high occurrence in a particular document. It is the product of TF and IDF.

$$TFIDF\ (t,\ d) = TF\ (t,\ d)\ *IDF(t)$$

TFIDF is a feature used regarding text classification. In addition, TF-IDF Vectors can also be determined at other level i.e., word-level and N-gram level, which is used

- ***Word level TF-IDF*:** Computes score for each solitary terms in various files.
- ***N-gram level TF-IDF:*** Calculates rating for the combination of Nterms collectively in various documents.

C. ALGORITHMS USED FOR CLASSIFICATION



Dataset of News Articles
Pre-processing of data (Tokenization, lemmatization, etc.,)
Feature extraction using TF,TFIDF
Training Classifier
Classification and categorization of news
FakeNews
Trusted accurate news

Fig 3. Process of Classification

1. Naïve-Bayes: This particular technique is used for classification defined on Bayes-theorem, it presupposes any occurrence associated with any functions in a class and is independently associated with the occurrence associated with other function or feature. It also states a way to estimate the posterior-probability.

$$P\ (c\ |x) = \frac{P(x|c).P(c)}{P(x)}$$

- P(c | x) →Posterior-probability of class given predictor
- P(x)→prior probability of class
- P(x | c) →likelihood (probability of predictor given class)
- P(x)→prior probability of predictor

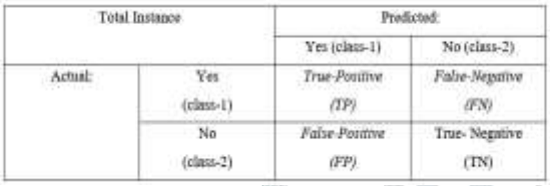
2. Passive Aggressive Classifier: these algorithms are online-learning algorithms. These algorithms remain inactive for any precise classification outcomes, transforms to aggressive with new miscalculation and improvisation and adaptation. Unlike, many other algorithms, it is not convergent. The main aim is to

update to correct the loss, triggering extremely minute difference in averaging the weight-vector.

D. METRICS TO MEASURE PERFORMANCE OF THE MODEL

Metrics is used to measures performance of the model in classifying or evaluating prediction.

1. Classification-Accuracy: This does not take frequent assessment metric for classification. It will define the amount of precise predication against the quantity of total forecasts. Though, this metric can give enough information to determine whether the design is a great one or even not it is insufficient.
2. Confusion-Matrix: This is also known as Error matrix, which will be a tabular representation that will show the particular performance of the model. It will be superior type of Backup table containing two dimensions- ‘Actual’ labelled on X-axis and ‘Predicted’ on Y-axis. The cells from the table are the particular quantity of predictions produced by the algorithm



Total Instance		Predicted:	
		Yes (class-1)	No (class-2)
Actual:	Yes (class-1)	*True-Positive* *(TP)*	*False-Negative* *(FN)*
	No (class-2)	*False-Positive* *(FP)*	True- Negative (TN)

Fig 4 Confusion-Matrix

True Positives: This accurately predicted positive certainty values.

True Negatives: It is accurately predicted negative certainty values.

False Positives: It is inaccurate prediction of negative values as positive values.

False Negatives: It is incorrect prediction of negative values as positive values.

3. Arrangement Rate or Accuracy: Scikit provides an ease report while functioning on classification difficulties which of the outputs are accurate, F1 report and support regarding each class.

Classification-Rate / Accuracy given by this relation:

$$\boldsymbol{Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}}$$

Nevertheless, there exists issues with accuracy and preciseness. The assumption is made as equal expenses for both types of errors. The 99% accuracy could be perfect, good, average, bad or worse depending upon the particular problem

Precision: This gives ratio between accurately predicted positive instances and total-predicted positive instances. Higher the precision means low False Positive rate.

$$\boldsymbol{Precision = \frac{T_P}{T_P + F_P}}$$

Recall (Sensitivity): Proportion of accurately predicted positive instances to the all of the instances in actual class - Yes.

$$\boldsymbol{Recall = \frac{T_P}{T_P + F_N}}$$

F1-Score: This gives weighted average of Precision and Recall. F1-score is generally more effective than accuracy, particularly in the cases where the distribution in class is uneven. Accuracy is best when false-positives and false-negatives have related instances. If the price of false-positives and false-negatives differs widely, then it is better to consider both Precision and Recall.

$$\boldsymbol{F1\ Score = 2 * \frac{(Recall * Precission)}{(Recall + Precision)}}$$

Let us study an instance, in which, there exist unlimited data-components of class-B and the only one component associated with class-A and the class will be predicted by model. ‘A’ against all the particular instances in the test data. Here,
Precision: 0.0
Recall: 1.0
Now,
Arithmetic-mean (AM): 0.5
Harmonic-mean (HM): 0.0

When considering the AM, it would have already 50% accurate. In spite of expecting the worst outcome. While we consider the HM, the F-measure is 0

N = 165	Predicted – ‘No’	Predicted – ‘Yes’
Actual – ‘No’	50	10
Actual – ‘Yes’	05	100

For that simplifying the confusion matrix, let’s have added, all the terms such as T_P, F_P, and so on and the row and column counts in subsequent table below:

N = 165	Predicted – ‘No’	Predicted – ‘Yes’	
Actual – ‘No’	50	10	60
Actual – ‘Yes’	05	100	105
	55	110	

Now,

Classification Rate/Accuracy:

Accuracy = $(T_P + T_N) / (T_P + T_N + F_P + F_N)$
= (100 + 50) / (100 + 5 + 10 + 50)
= 0.90

Recall:

Recall allows us to predict when is it an actual yes and the frequency of predicted yes.

Recall = $T_P / (T_P + F_N)$
= 100 / (100 + 5)
= 0.95

Precision:

Precision allows us to know when the module predicts yes and how frequently is it accurate.

Precision = $T_P / (T_P + F_P)$
=100/ (100+10)
= 0.91

F-measure:

Fmeasure = (2 * Recall * Precision) / (Recall + Precision)
= (2 * 0.95 * 0.91) / (0.91 + 0.95)
= 0.92

V. EXPIREMENT, RESULTS AND ANALYSIS

Experiment was performed using the above algorithms applying Vector features- Count Vectors and Tf-Idf vectors at word level and Ngram-level. Accuracy was observed for all models. Kfold cross-validation approach is used to increase the performance of the presented designs. In the first phase of our experiment, text classification has been used on the article's body in two different publicly available datasets.

In the next phase, Experiment has been performed on the responses collected about a pair of Fake Information and Real information claims removed from Twitter and fb.

A. DATASET SPLIT USING K-FOLD CROSS VALIDATION

This cross-validation technique was used for splitting the dataset randomly in to k-folds. (k-1) folds were used with regard to build the model while kth fold was used to examine the effectiveness associated with the model.

B. SET OF EXPERIMENTS CONDUCTED

1. Experiment (proposed model):

The responses were grouped using count vector and tf-idf vector at two levels:

Word Level – single word was chosen for this experiment.

Classification Accuracy at Word Level:

Accuracy	**Multinomial Naïve Bayes**	**Passive Aggressive Classifier**
Using Count Vector	89.34	83.3
Using Tf-Idf vector	85.65	93.83

N-gram Level –the range of n-gram from 1 to 3 has been kept by choosing from one word to at most 3 words which was considered and experiment was performed.

Classification Accuracy at N-gram Level:

Accuracy	**Multinomial Naïve Bayes**	**Passive Aggressive Classifier**
Using Count Vector	86.4	81.8
Using Tf-Idf vector	77.3	90.9

Classification Accuracy at word level performed much better than N-gram level as we can see through the above tables. The accuracy with regard to Multinomial Naïve Bayes with Tf-Idf at N-gram level was the lowest at 77.3% while Passive Aggressive Classifier, using Tf-Idf vectors carried out well at each level and the accuracy was over 90%

2. Classification Reports:

PAC-TFIDF:

Classification Report	Precision	Recall	F1-score	Support
Fake	92	86	89	1008
Real	88	93	90	1083

3. CONFUSION MATRIX:

➢ **MNB- TFIDF**

VI. T_P=739, T_N=1052, F_P=31, F_N=269

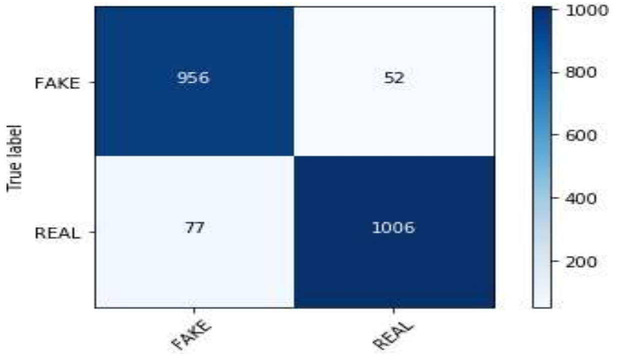


Fig 5. Confusion Matrix for MNB-TFIDF, Split 1

- TP=865, TN=1003, FP=80, FN=143

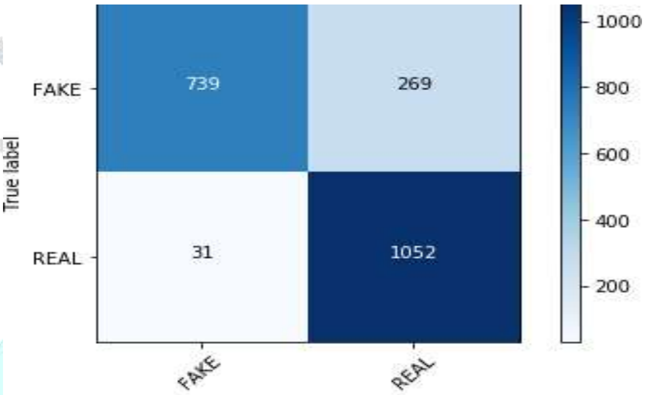


Fig 6. Confusion Matrix for MNB-TFIDF, Split 2

➢ PAC- TFIDF

- T_P=956, T_N=1006, F_P=77, F_N=52

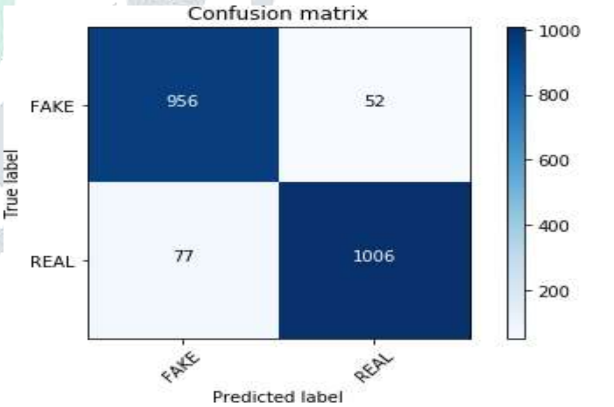


Fig 7. Confusion Matrix for PAC-TFIDF

VI. CONCLUSION

User's viewpoint on online or social-media platform posts can be well applied to determine the veracity of information. Dissemination of Fake Information on social-media is fast and therefore the proposed methodology, can serve as a basic building block for Fake-information-detection. With highest classification accuracy of 93. 2%, sensitivity of 92% and count and Tf-Idf vector is a much better model as compared to others. By adding large amount of data to the dataset can test the consistency of the model and performance as well. In addition, gathering real information that almost appears as Fake Information will enhance the training of the model. More linguistic based features can be applied on responses to determine the information truthfulness. Social-media has a very significant role in the verification process. The move from traditional mass media to social mass media and fast diffusion of information, bank checks this limitation. As a result, by exploring more social media features inside our experiments, and combining them we can produce an efficient and reliable system for detecting Fake Information.

VII. REFERENCES

[1] www.geeksforgeeks.org

[2] upcommons.upc.edu

[3] en.wikipedia.org

[4] Desislava Ivanova, Plamenka Borovska. "Scalable framework for adaptive in-silico knowledge discovery and decision-making out of genomic big data", AIP Publishing, 2018 Publication

[5] Houda Benbrahim. "A Fuzzy Semi-Supervised Support Vector Machines Approach to Hypertext Categorization", IFIP – The International

[6] Submitted to Visvesvaraya Technological University, Belagavi

[7] Divya Khanna, Prashant Singh Rana. "Improvement in prediction of antigenic epitopes using stacked generalisation: an ensemble approach", IET Systems Biology, 2020