Indian Institute Of Information Technology
Allahabad

7th Semester Mini Project

Fake News Detection

Submitted To:
Dr. K.P.Singh

Submitted By:
Pallavjeet Singh Nirwan (IIT2015131)
Vaibhav Bansal (IIT2015138)
Sai Akhil Aloor (IIT2015147)

CERTIFICATE

I hereby recommend that the mini project report prepared under my supervision, titled **"Fake News Detection"** be accepted in the partial fulfillment of the requirements of the completion of mid-semester of seventh semester of Bachelor of Technology in Information Technology.

Date: November 19, 2018
Place: Allahabad

Supervisor :
Dr. K.P.Singh

Contents

1 Introduction

Fake news is news which are created intentionally to misguide the readers. It is a type of propaganda which is published in the form of genuine news. Fake news is spread through traditional news media and social media [1]. Fake news has been a problem from a long time. With the introduction of social media, the spread of fake news is increased and it became difficult to differentiate between true news and fake news. The spread of fake news is a matter of concern as it manipulates the public opinions. During the American Presidential elections of 2016, it was estimated that over 1 million tweets are related to fake news "Pizzagate" by the end of the elections. The wide spread of fake news can have a huge negative impact on individuals and society as a whole.

We developed a model which accurately determines whether an article is fake or not using machine learning and NLP techniques.

2 Motivation

The extensive spread of fake news can have a serious negative impact on individuals and society. It has brought down the authenticity of news ecosystem as it is even more widely spread on social media than most popular authentic news. It is one of the biggest problems which has the ability to change opinions and influence decisions and interrupts the way in which people responds to real news. It have political influence, can encourages mistrust in legitimate media outlet, influence financial markets, damage to individual's reputation.

During the American presidential elections of 2016, a survey revealed that many young men and teens in Veles were running hundreds of websites which published many false viral stories that supported Trump. These fake news influenced many people which affected the election results. This is just a small instance of how the spread of fake news can influence people.

Many organizations have come forward to stop the spread of fake news. Eg. Google app uses Artificial Intelligence to select stories and stop fake news.

We aim to develop a model using machine learning and NLP techniques to determine whether a news is fake or real.

3 Problem Definition

Fake news is a real menace as it can quickly spread panic among the public. It can also affect major world events, as was seen in the US Presidential Elections. With the flood of news arising from online content generators, as well as various formats and genres, it is impossible to verify news using traditional fact checkers and vetting. To tackle this problem of quick and accurate classification of news as fake or authentic, we provide a computational tool.

4 Objectives

The major objectives of this project are:

1. Taking the help of linguistic cues to develop a machine learning based model for accurately determining whether the given news is fake or authentic.

2. To get high accuracy to determine a news is fake or true.

5 Literature Review:

1. In the research paper[2] "Syntactic Stylometry for Deception Detection" by Feng, Song and Banerjee, Ritwik and Choi, Yejin, they focus on the language patterns followed by deceptors in their language. Most liars use their language strategically to avoid being caught, in spite of their attempt to control what they are saying, language “leakage” occurs with certain verbal aspects that are hard to monitor manually such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage. The research focuses on user reviews and essays, but is equally applicable in fake news detection as well, where the author tries to deceive the readers. For detecting these linguistic patterns, the research uses shallow and deep syntax analysis which use POS(parts-of-speech) tags and Probabilistic Context Free Grammars(PCFG) respectively.

2. In their research paper [3] "Automatic deception detection for detecting fake news" by Conroy, Niall J. and Rubin, Victoria L. and Chen, Yimin, they discussed two methods for Fake News detection. The first one is linguistic approach, this discusses the various syntactical and semantical features that are useful in deception detection. It uses deep syntax analysis with context free grammar generation using stanford parser. The basis of semantic analysis provided in this research is that, the author may use contradictions and omit facts while writing, whereas a writer of a truthful review is more likely to

make similar comments about aspects of the product as other truthful reviewers. Network approach depends on querying existing knowledge networks, or publicly available structured data, such as DBpedia ontology, or the Google Relation Extraction Corpus (GREC). It also takes into account that on social media, the authentication of identity of the user posting an article is paramountal for the notion of trust.

3. In their research paper [4] by Veronica P ´ erez-Rosas , Bennett Kleinberg , Alexandra Lefevre Rada Mihalcea1, thet focused on the automatic identification of fake content in the online news. They introduced two novel datasets for the task of fake news detection, covering seven different news domains. They build the fake news detection models by extracting several linguistic features like n-grams, punctuations, pyscholinguistic features,readibility, syntax etc. They used LIWC to generate these features. Their best models achieved accuracies which are similar to the human ability to spot fake news.

4. In their research paper [5] "Evaluating Machine Learning Algorithms for Fake News Detection" by Shlok Gilda, he explored various Natural Language Processing techniques for the detection of fake news detection. He applied techniques like TF-IDF of bi-grams and probabilistic context free grammar (PCFG) detection on a data corpus of about 11,000 articles. He tested the dataset on multiple classification algorithms like SVM, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees, and Random Forests. He found that the Stochastic Gradient Descent model gives an accuracy of 77.2%.

5. In their research paper [1] "Fake News Detection on Social Media : A Data Mining Perspective" by Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, they present a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics and representative datasets. For feature extraction they used : News Content Features - Linguist and Visual based. Social Context Features - It include features from users, posts and network. and for Model Constructions they used: News Content Models - knowledge based and style based. Social Context Models - Stance and propagation based.

6. In their research paper [6] "Automatic detection of deception in child-produced speech using syntactic complexity features" by Maria Yancheva, Frank Rudzicz, the evaluations are done using a novel set of syntactic features, including measures of complexity. Their results show that sentence length, the mean number of clauses per utterance, and the Stajner Mitkov measure of complexity are highly informative syntactic features.

6 Datasets:

We have 3 datasets for fake news detection, we will use the required attributes of these datasets to train our model.

- Getting Real about Fake News Dataset- Kaggle
- Fake News Detection Dataset- Kaggle
- Fake News Dataset - Kaggle

The total datasets have about 37,808 data points. But after preprocessing the data, (i.e) removing duplicates and removing data points which are NULL, there are 27865 data points out of which 15343 articles are REAL and 12522 articles are FAKE.

The distribution of data points in the dataset is as follows:

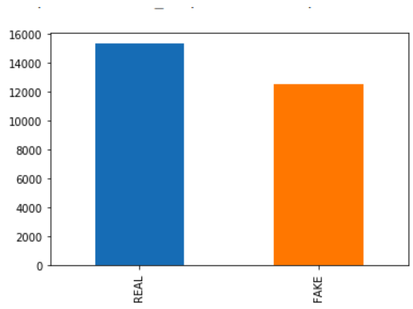


Figure 1: Distribution of dataset

7 Methodology

The approach proposed for this project is:

1. Data Preprocessing
2. Generating News Feature Vector
3. Classification

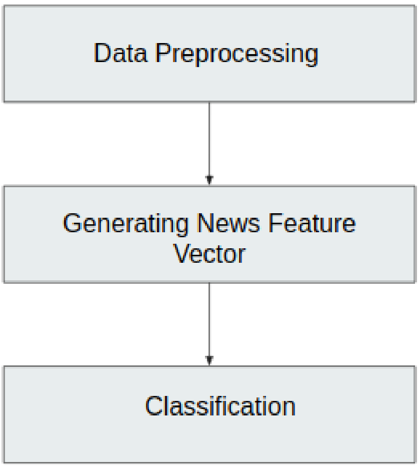


Figure 2: Flowchart

7.1 Data Preprocessing:

Data preprocessing is done to convert the raw data into a required format. Data preprocessing can be done by various methods like data cleaning, data reduction, data integration etc. In this project, the datasets are collected from different resources which have different formats and attributes. Hence, the data can be duplicate and they may contain some attributes which are not useful. So, we convert the data into our required format with required attributes which are used to train our model.

7.2 Generating News Feature Vector:

The most important part of detecting if a given news is fake or not is to convert the news article into a news vector which contains the important features which are used to determine the nature of the news.

There are several ways to generate this feature vector[3] [4] . We tried different approaches for the same to determine which method gives the best accuracy. Some of the methods are:

7.2.1 Bag of Words:

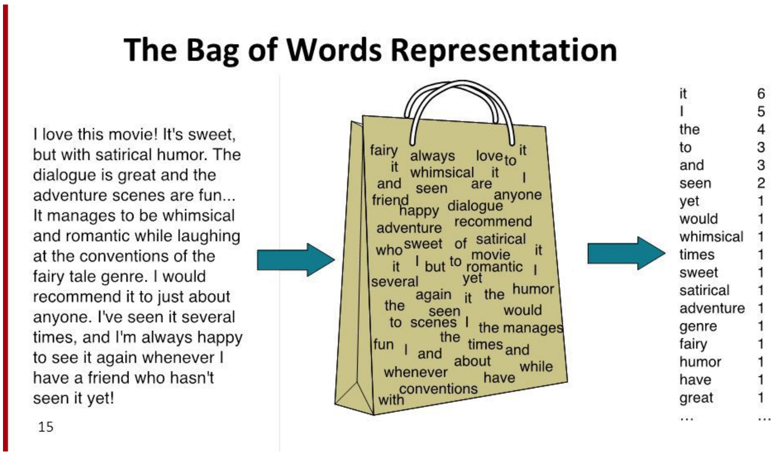
Bag of words is a way of representing text in a format which can be easily processed by the machine learning algorithms. BoW is one of the ways of extracting features from text. In this type of text representation, majorly 2 things are involved:

- Vocabulary of known words.

- Measure of the presence of known words

In BoW, representation, the order of words or structure of the sentence is not considered, it is only concerned with whether the word is present in the document or not. The BoW is carried out as follows:

- Collect all the unique words in all the documents.
- Now, represent the documents in a vector of all the unique words by counting the number of times each word appears in that document.



The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Figure 3: Example of Bag of Words

There is another type of BoW named binary BoW, in which instead of count of the count of each word, it just checks whether a word is in that document and then represent with a 0 or 1.

7.2.2 TF-IDF:

TF-IDF stands for term frequency-inverse document frequency.TF-IDF is a method used to represent text in a format which can be easily processed by the machine learning algorithms. It is a numerical statistic that shows how important a word is to a document in a word corpus. The importance of a word is proportional to the number of times the word appears in the document but inversely proportional to the the number of times the word appears in the corpus. The tf-idf weight is composed of 2 terms:

1. **TF (Term Frequency):** Term frequency is defined as the frequency of a word in the document. TF is calculated as:
   TF(w) = (Number of times word 'w' appears in the document) / (Total number of words in the document).

2. **IDF (Inverse Document Frequency):** This measures how important a word is in the document. For example, words like and, of, the, a appears lot of times but they are less important. Thus most repeated terms are given less weights and less frequent terms are given more weights. IDF is calculated as: IDF(w) = log_e(Total number of documents / Number of documents with word ‘w’ in it).

The TF-IDF weight is given to each word by calculating TF*IDF values.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 4: Formula for TF-IDF

For generating the news vector, we calculate the tf-idf values of the bigrams and represent the tf-idf vector of that bigrams. We choose bigrams over unigrams because it gives the context.

n-grams: n-gram is a contiguous sequence of n terms from a given text. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram", size 3 as a "trigram" and so on. with larger n, a model can store more context.

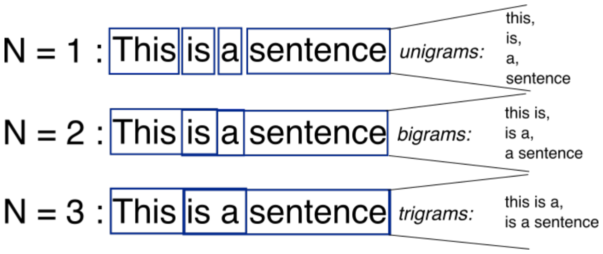


Figure 5: Example of n-grams

7.2.3 Shallow and Deep syntactical Analysis:

We generated POS(part-of-speech) tags using the Spacy library. Our POS features will be encoded as tf-idf values for each for these tags. Research paper by Feng et al [2] states that, even though POS tags are effective in detecting fake product reviews, they are not as effective as words. Therefore, we strengthen POS features with unigram/bigram features.

For deep syntactical analysis we used the Stanford/Berkeley parser to generate CFG rules for the sentences and we encoded these rules with tf-idf values for each production rule.

7.2.4 Semantic Analysis:

A widely used open-source resource for incorporating semantic information is Empath(developed by Stanford). Empath is a lexicon of words grouped into semantic categories relevant to psychological processes. Several research works have relied on semantic analysis to build deception models using machine learning approaches and showed that the use of semantic information is helpful for the automatic identification of deceit. Empath has 194 semantic categories, some of these semantic classes are emotional tone(positive or negative), anger, nervousness.

We get a score between 0-100 for each semantic class. The lexicon we get is converted to a TF-IDF vector by taking the score for a semantic class(like nervousness) as its frequency.

7.2.5 Combining features to form final news vector:

We considered 3 methods for generating feature vectors:

1. TF-IDF bigram vector of the news article.

2. Feature Vector generated by Syntax Analysis of the news article.

3. Feature Vector generated by the semantic analysis of the news article.

After generating these features and generating their individual feature vector, we have to combine these features to form the final news vector on which classification is performed.

The method we approached for combining the feature vectors is

1. Take the most important features for the 3 feature vectors.

2. Assign weights to each vector and then take the weighted combination of the 3 feature vectors to generate the final feature vector. If x is the weight corresponding to the first feature vector, y for the second, and 1-x-y for the third. The final feature vector will be the linear combination of these feature vectors multiplied by their corresponding weights.

7.3 Classification:

After generating the news feature vector, now we classify the vector to whether it is fake or real. We aim to use the following classification algorithms[5] for the purpose of classification:

7.3.1 Naive Bayes:

Naive bayes is a supervised learning algorithm which is used for classification. It is based on bayes theorem assuming that features are independent of each other. It calculates the probability of every class, the class with maximum probability is chosen as the output.

7.3.2 Random Forests:

Random Forests is a bagging type of ensemble model in which the base model for bagging is decision tree. In addition to bagging (i.e taking random samples of total data and train those samples independently and then take the majority voting of numerous samples of the total dataset to find the output of the total dataset), there is feature bagging (i.e column sampling in which not all the columns/features are taken into consideration while training but rather random samples of features are considered while training different samples.

$$\text{Random Forest} = \text{Bagging with decision tree as base model} + \text{feature bagging}$$

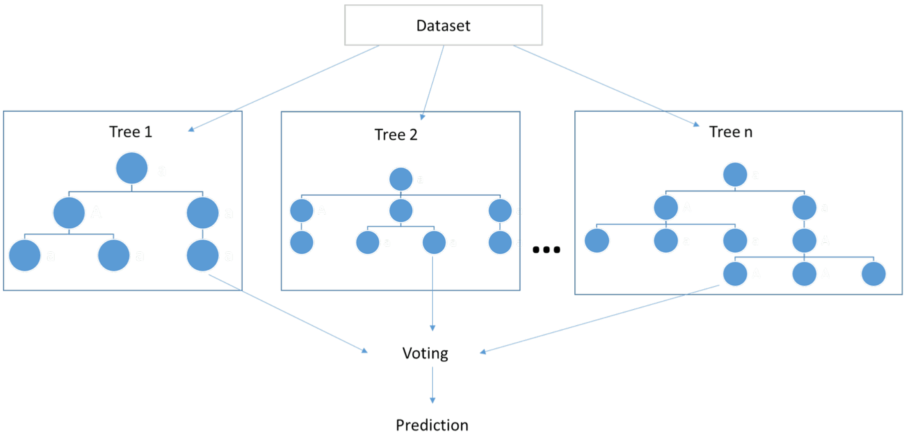


Figure 6: Example of Random Forests

7.3.3 Gradient Boosting:

Gradient boosting is an example of boosting ensemble model. Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. Boosting tries to convert a weak learner to become better. It learns from its mistakes/error and tries to reduce the error.

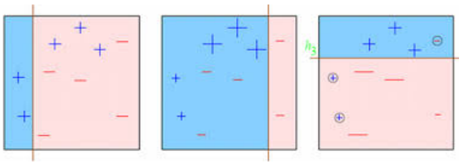


Figure 7: Example of Gradient Boosting

8 Results:

After generating the feature vectors they were combined with different weights. The results are compiled int the following table :

Weights of the feature vectors and corresponding results					
Bigrams vector	Syntax vector	Semantic vector	Multinomial Naive Bayes	Random Forests	Gradient Boosting
1	0	0	82.9%	84.7%	86.5%
0	1	0	66.5%	86.6%	88.3%
0	0	1	54.6%	67.0%	69.8%
0.33	0.33	0.33	91.2%	86.0%	91.4%
0.5	0.5	0	92.2%	88.4%	91.4%
0.5	0	0.5	90.5%	80.3%	85.8%
0	0.5	0.5	89.4%	79.9%	88.3%
0.2	0.2	0.6	87.8%	86.2%	90.8%
0.2	0.4	0.4	89.8%	85.5%	90.8%
0.2	0.6	0.2	90.9%	85.6%	90.9%
0.35	**0.5**	**0.15**	**92.7%**	**89.7%**	**91.5%**
0.4	0.2	0.4	91.1%	85.0%	91.3%
0.4	0.4	0.2	92.1%	86.3%	91.4%
0.6	0.2	0.2	91.6%	86.5%	91.3%
0.4	0.5	0.1	92.3%	87.3%	91.4%

The best accuracy we achieved is 92.7% corresponding to the weights (0.35,0.5,0.15) for the feature vectors.

Confusion matrix corresponding to the weights (0.333,0.333,0.333) (equal weightage) :

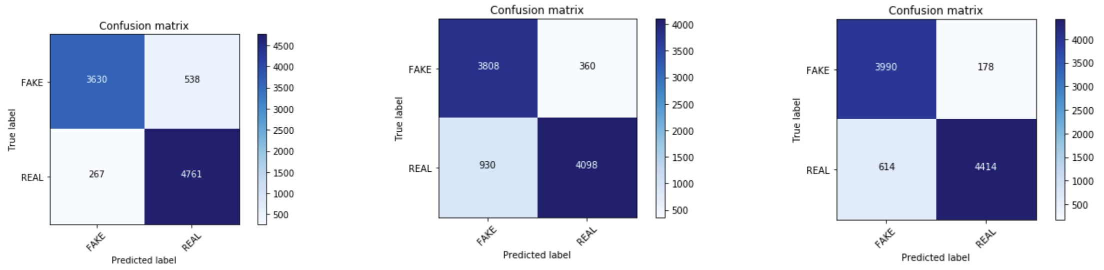


Figure 8: Confusion matrices for Naive Bayes, Random Forests and Gradient Boosting respectively

Confusion matrix corresponding to the weights (0.35, 0.5, 0.15) which gave maximum accuracy is:

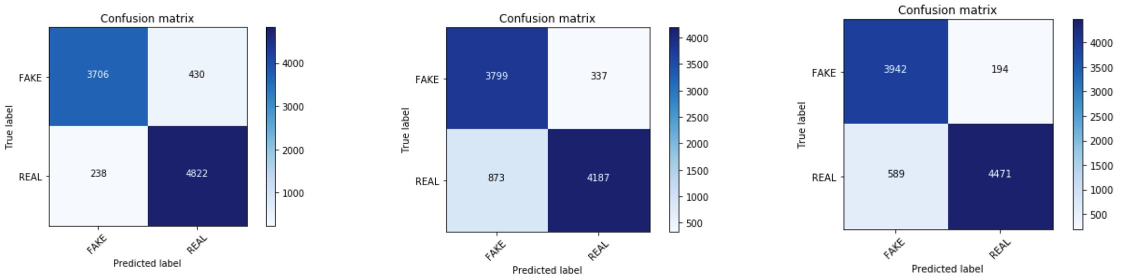


Figure 9: Confusion matrices for Naive Bayes, Random Forests and Gradient Boosting respectively

9 Conclusion:

We observed that all the 3 features are paramount in detecting fake news when combined together. We achieved the best result with an accuracy of 92.7% by using the weights (0.35, 0.5, 0.15) for feature vectors derived by bigrams, syntax and semantic analysis. Thus we conclude that linguistic features are pivotal in detecting whether an article is real or fake.

10 Future Work:

1. Our reseaarch focuses on daily news articles which have on average around 1000 words. It is difficult to detect linguistic cues in single(or few) statement news. Some other method can be researched upon for these cases.

2. For designing a fake news detector for social media like Facebook or twitter, we can take into account the user information, user authenticity and origin of the news.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *CoRR*, vol. abs/1708.01967, 2017.

[2] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, (Stroudsburg, PA, USA), pp. 171–175, Association for Computational Linguistics, 2012.

[3] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, (Silver Springs, MD, USA), pp. 82:1–82:4, American Society for Information Science, 2015.

[4] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *CoRR*, vol. abs/1708.07104, 2017.

[5] S. Gilda, "Evaluating machine learning algorithms for fake news detection," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pp. 110–115, Dec 2017.

[6] M. Yancheva and F. Rudzicz, "Automatic detection of deception in child-produced speech using syntactic complexity features," in *ACL*, 2013.