National College of Ireland

Hybrid feature extraction method using Latent Dirichlet Allocation and Term Frequency Inverse Document Frequency to detect fake news

MSc Research Project
Data Analytics

Zainul Abedin Khatik
Student ID: x18110096

School of Computing
National College of Ireland

Supervisor: Ms. Sidra Bashir

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Zainul Abedin Khatik
Student ID:	x18110096
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Ms. Sidra Bashir
Submission Due Date:	20/12/2018
Project Title:	Hybrid feature extraction method using Latent Dirichlet Allocation and Term Frequency Inverse Document Frequency to detect fake news
Word Count:	6446
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid feature extraction method using Latent Dirichlet Allocation and Term Frequency Inverse Document Frequency to detect fake news

Zainul Abedin Khatik
x18110096

Abstract

With the rapid growth of online newspaper people get updates on their fingertip. It is not difficult to generate fake content through social media and due to this it leads to a large volume of content to analyze. This information available online is extremely diverse which gets difficult to cross verify each and every article manually by the editors. For humans it is difficult to detect when fake news from real as the news is written in same style. In this research, we proposed a hybrid feature extraction technique using TF-IDF and LDA along with machine learning algorithms to build an automatic fake news detection model. To evaluate the performance different machine learning algortihms such as Linear SVM, Radial SVM, Random Forest and Naive Bayes is employed. In our proposed hybrid method, first, weights are assigned to the features using Term Frequency- Inverse Document Frequency. Then, top 50 terms are selected based on their weights, this top 50 features acts as an input to LDA. Futher, LDA is applied on the extracted features by TFIDF. Later, these extracted features from TFIDF and LDA will be analysed using different supervised classification models. Futhermore, the outcome of our experiments was that although our proposed hybrid method was able to perform better than the baseline classifier. But, Random Forest (76.07% accuracy) works the best with the features extracted by TF-IDF model individually in the process of fake news classification. For evaluation we have used Accuracy, Precision, Recall and F1.

1 Introduction

Gone are the days when people used to wait for the newspaper every morning. However, with the rapid growth of online newspaper people get updates on their fingertips. Also, the recent growth of online social media has vastly facilitated the way people communicate with one another. People use social media to connect with each other, share information and to be aware of the trending events. However, it cannot be ignored that the recent information that appears in social media can sometimes be misleading and fake. This can lead to serious problems in the society, one being the 2016 U.S presidential election campaign which was majorly influenced by fake news. During the election fake websites had produced false content, which had generated around 8.7 millions shares, comments and reaction on Facebook. Ironically, which was larger than the total of 7.3 millions news

stories published by Buzzfeed, well-known news website[1]

With the increase in fake news, it becomes critical to detect when news is false. It is easier to generate fake content through social media and due to this it leads to a large volume of content to analyze. This information available online is extremely diverse which leads to large number of subjects and increases the complexity.

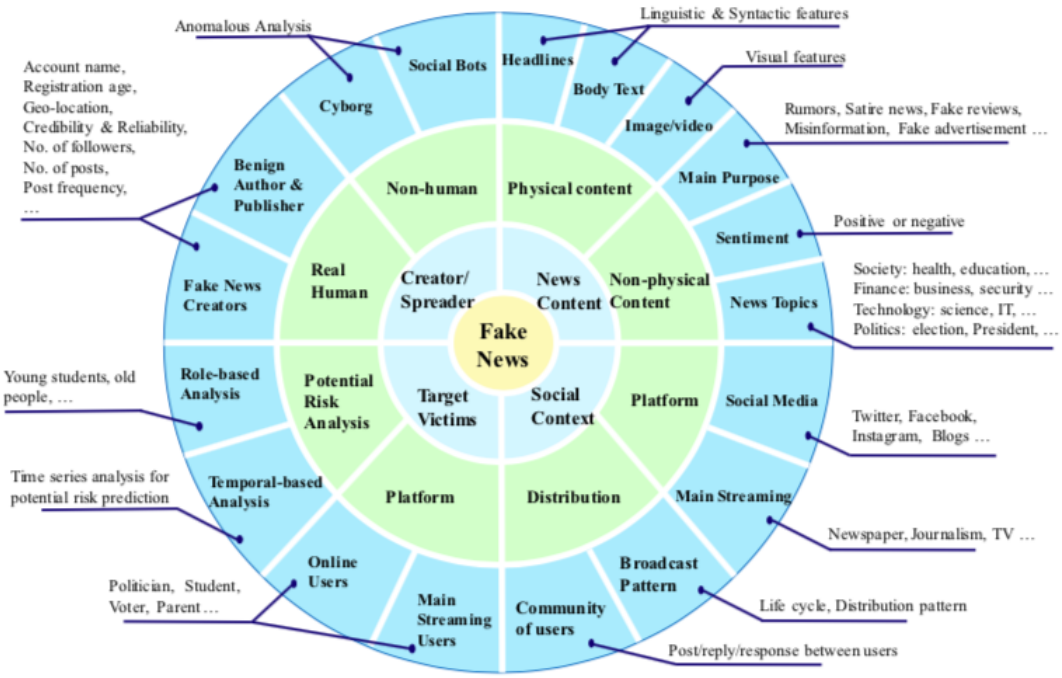


Figure 1: Example of a Facebook user sharing fake news. Adapted from: (Zhang and Ghorbani; 2019)

The Figure 1 shows the variety of online fake news available, onion shaped graph contains fake news in the centre and 4 types around it that are News Content, Social Context, Target Victims and Creator/Spreader.

1. News content is the body of the article.
2. Social Context is how the news is spread across the internet.
3. Target Victims are people who have been the victim of online fake content, mostly they are senior citizens, students and so on.
4. Creator or Spreader maybe humans or bots, where humans who do not have any knowledge about the subject might unintentionally spread fake news and there users who are paid to make false content on purpose for political gains or personal gains.

1.1 Background and Motivation

Currently, there are two methods to detect fake news, manually and automatic. Fact checking organisation such as Polifact, Classify.new, FactCheck.org, TruthOrFriction, Factmata, Snopes, etc are tackling fake news manually. They are using traditional methods to cross verify the news where reporters evaluate each and every news article manually.

[1]https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

However, due to vast amount of news published daily, it is difficult and time-consuming process to go through every article which has been published. To solve this issue, many researchers have a build a proof of concept to automate the process with the help of machine learning algorithms for fake news detection.

Researchers Jain and Kasbe (2018), Granik and Mesyura (2017) are have used traditional machine learning algorithms to solve fake news detection. However, Ahuja et al. (2019), Kotteti et al. (2018) and Xu et al. (2018) observed that extracting features before classification phase, improvises accuracy of machine learning algorithms. In text classification, learning from high dimensional data is a challenging task. As the documents contains enormous amount of information such as phrases, words, etc., this leads to high computation burden in the learning phase. Additionally, features that are not relevant can harm the accuracy of the classification models. Thus, there is a need of feature extraction to reduce the text size. Number of features will get reduce after applying feature extraction technique on the data, this will prevent over-fitting and the training time to train will be less. In this research, two feature extraction techniques are employed such as Term frequency-inverse document frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) on the of news articles.

Term frequency-inverse document frequency assigns weight according to the importance of the word to a document in the document set. It means that rarer the term, higher is the TF-IDF weight. So, the words such as "this", "what" and "if" occur very often are given a low score since they do not much significant impact on the document. Researchers Xu et al. (2018), Agudelo et al. (2018) have employed TF-IDF as a feature extraction method to improvise the accuracy of machine learning algorithms. However, as TFIDF is a bag of words model, it does not capture co-occurrence and semantics. To solve this issue of polysemy and synonymy, Blei et al. (2003) recommended LDA as an alternative method for feature extraction.

LDA is a generative document model which is capable of topic modelling as well as dimensionality reduction (Blei et al.; 2003). Without any prior knowledge, LDA explores semantic information among words and extracts underlying topics from a corpus. It creates topic per document and words per topics, modelled as Dirichlet distributions. Few researchers have used Latent Dirichlet Allocation model for feature extraction. For example, Prihatini et al. (2018) have used LDA for feature extraction on Indonesian text. However, LDA does not consider the importance of a word in a document. Thus, by employing TF-IDF term weighting scheme words are assigned weight to according to its importance.

On a whole, taking influence from Xue (2019) where researchers have used a combination of LDA and Word2vec for feature extraction. Likewise, in this research, a hybrid feature extraction method is proposed using TFIDF and LDA, in an attempt to improvise the accuracy of machine learning algorithms to detect fake news from real. In this research, three experiments are being performed. The first and the second are performed by employing TFIDF and LDA individually along with machine learning algorithms. Lastly, using proposed hybrid feature extraction method of TFIDF and LDA to extract features from real world news articles along with classifiers.

1.2 Research Question

Would the combination of Term Frequency - Inverse Document Frequency(TF-IDF) and Latent Dirichlet Allocation (LDA) for feature extraction improve the accuracy in detecting fake news compared to features extracted individually by TFIDF and LDA?

1.3 Objectives

Objective 1: Discuss state of art results in the area of fake news detection.
Objective 2: Extract features by using TFIDF and LDA along with machine learning algorithms.
Objective 3: For comparison purpose, extract features using TFIDF and LDa individually along with machine learning algorithms.

The rest of sections are as follows: Section 2 covers work done researchers in the area of fake news detection. Section 3 covers methodology of the research project. Section 4 covers design of the project and explaination of TFIDF and LDA models. Section 5 covers implementation. Section 6 covers evaluation and section 7 covers discussion. And lastly, Section 8 covers conclusion and future work.

2 Related Work

This sections represents work done by researchers in the field of fake news detection.

2.1 Dectecting fake news using data mining techniques

Granik and Mesyura (2017) A study was conducted by Granik and Mesyura (2017) to detect fake news article which are posted on public social media platform Facebook. The authors of the study have used Buzzfeed data set which consists of Facebook posts. On this dataset, the analysis has been done by considering individual post of Facebook as a news article. BuzzFeed dataset consists of four types of articles named mostly true, mostly false, a mixture of true and false and no factual content. It contains of 2200 posts or articles. But, Granik and Mesyura (2017) have considered only two factors of data for their analysis. They are, mostly true and mostly false. In pre-processing and filtration process of data only 1700 articles where remining for data training and testing. Authors have employed Naïve Bayes classification model for their analysis and achieved an accuracy of 76% in detecting fake articles of Facebook. Similarly, Jain and Kasbe (2018) have conducted a study to detect fake articles posted on Facebook. The data for analysis is taken from GitHub which contains of 11000 news articles. The GitHub dataset consists of data with combination of fake and real news and divided into four categories. The columns of the dataset are index, title, label and text which consists of news articles from science and technology, entertainment, business and health categories. When compared with Granik and Mesyura (2017) paper, the authors Jain and Kasbe (2018) have introduced a bag of words model for generating vocabulary. The same classification algorithm is used by Jain and Kasbe (2018) and achieved an AUC score for the title and text is 0.806 and 0.912 in detecting fake news from Facebook. The issues with the bag of model like word order is solved in Jain and Kasbe (2018) paper by introducing n-grams. The results where improved to 0.807 for title and 0.931 for text by solving issue.

In another research study, Fontanarava, Pasi and Viviani (2017) have taken dataset called Yelp review (YelpZip and YelpNYC). The dataset is divided into reviews, reviewers and entities. The authors were able to achieve 77.6 precision, 86.1 recall, 81.6 f-score and 81% accuracy by employing random forest algorithm. These results achieved by Fontanarava et al. (2017) is much better than that of results got by Mukherjee et al. (2013) because the authors have only focused on reviews rather than all entities in the dataset, where authors have used same dataset Yelp review with random forest classification model. The results got by Mukherjee et al. (2013) are 71.6 precision, 66.4 recall, 68.9 f-score and 70% accuracy.

Bourgonje, Moreno Schneider and Rehm (2017) Ferreira and Vlachos (2016) (Srivastava, Rehm and Moreno Schneider; 2017).

(Kotteti et al.; 2018). In section 2.2 Kotteti, Dong, Li and Qian (2018), Hardalov et al. (2016), Xu et al. (2018), Gilda (2017) and Ahmed et al. (2017)

As the data is in high dimension, it becomes a necessity to reduce the dimensionality of the data. In section 2.2 and 2.3 researchers demonstrated that how extracting features can improvise the performance of machine learning algorithms.

2.2 Feature extraction using TF-IDF

To improve accuracy of machine learning algorithm in detecting fake news Kotteti, Dong, Li and Qian (2018) proposed a novel data preprocessing technique to handle missing values by using data imputation method for numerical and categorical features. The authors have employed TFIDF to extract features from the news content. For classification purpose SVM, DT, Multilayer Perceptron and Gradient Boosting algorithm are used. The dataset used is LIAR dataset which is human labelled and contains around 12000 records which is evaluated by PolitiFact editors. There are six target variables pants-fire, true, false, barely-true, mostly true, half true. The dataset has a combination of both categorical and numerical features. The authors have considered Wang (2017) model as a baseline for their study. Wang (2017) model is build using hybrid CNN by integrating metadata with text to detect fake news. Comparing the work done by Kotteti et al. (2018) and (Wang, 2017), shows that the experimental results of Kotteti et al. (2018) with Multilayer Perceptron classifier was able to improve the accuracies on test set and validation set by 16% and 21% respectively as compare to Wang (2017) model. Wang (2017) model might have got better results if the authors had extracted relevant features before feeding to the classifiers. On a whole, it indicates that despite Wang (2017) employing deep learning techniques, Kotteti et al. (2018) were able to get better results by using TFIDF.

In another research, Xu et al. (2018) proposed a threefold method to detect fake news. Firstly, authors analysed an interesting fact that how fake news gets disappeared after a certain period of time from the websites. So, the first step of the threefold method is to check the reputation of the publishers based on their registration, since when the website is being operated from, and how quickly it disappears from the internet. The second step is to check the similarity of fake and real news based on the important terms via TFIDF and LDA topic modelling. Lastly, the third step to check document similarity by using Jaccard similarity to detect, predict and classify fake and real news. The authors here have experimented with top ten terms most important words produced by TFIDF and top 3 topics with top 5 words from LDA. The experimental result shows that just by using TFIDF and LDA topic modelling, it is difficult to classify fake from real. However,

document similarity of real news in training set is similar to real news in test set.

Agudelo, Parra and Velandia (2018) implemented two feature extraction technique countvectorizer and tfidfvectorizer. The dataset contains around 10558 news articles. For classification purpose the author has taken Naïve Bayes algorithm to classify real and fake news. The experimental results shows that the model is able to predict with 88% accuracy. Likewise, compared to study conducted by Chiu et al. (2013) , the authors implemented ngrams and LDA topic modelling for semantic purpose. For classification purpose, the authors have used SVM as a classifier and were able to achieve accuracy of 95%. By comparing experimental results of both Agudelo et al. (2018) and Chiu et al. (2013), it shows that how semantics can play an important role.

In another research, Ahuja et al. (2019) analysed the impact of feature extraction techniques on sentimental analysis from twitter data. The authors have focused on only TF-IDF and N-gram (value of n=2) on the performance of sentimental analysis and made a comparative study between them. The authors applied TF-IDF and N-gram individually on six classification algorithms (Decision Trees, SVM, KNN, Random Forest, Logistic Regression and Naïve Bayes) to learn which features are better. The S-S Tweet dataset used here is from twitter, there are 4242 tweets in which there are 1251 are positive tweets, 1953 are neutral tweets, and 1037 are negative tweets. The authors found out that Logistic regression performed better in both the cases and TF-IDF gave better results by 3% to 4% as compare to N-gram features. The authors here only took TF-IDF and N-grams into consideration, but did not take other feature extraction technique such as LDA, Word2vec, etc. From the paper it can be noticed that how feature extraction techniques help to improve the performance of machine learning model. On a whole, TF-IDF performed better than N-gram by 3% to 4%.

Gilda (2017) proposed a technique to evaluate the performance of machine learning models with feature extraction method such as TFIDF using bigrams and PCFG (probabilistic context free grammar). The authors have obtained dataset from Signal media which has 11,000 articles and for classification purpose the authors have used SVM, Stochastic Gradient Descent, Bounded Decision Trees, Gradient Boosting and Random Forest. For evaluation of results authors have used precision, accuracy and recall and the authors have made Naïve Bayes (with accuracy 69%, precision 54% and 54% recall). The authors have made Random forest (with 32% accuracy, 32% precision and 56% recall) as their baseline model without applying any feature extraction technique on the dataset. To evaluate the performance of the models the authors have implemented TFIDF, PCFG and combination of both for feature extraction and then fed to the classifiers. The best results obtained from the experiment are by Stochastic Gradient Descent using TFIDF bigrams with an accuracy of 77%, AUC 88%, recall 45% and precision 88%. This explains how extraction features can increase the performance of the classifers.

Ahmed, Traore and Saad (2017) conducted a research to analyse the impact of TFIDF on performance of machine learning algorithms. The authors have used TFIDF and Term Frequency(TF) for feature extraction techniques with Ngrams (from n=1 to n=4) to detect fake news. The authors have used 12600 articles fake news dataset from Kaggle and combined it with 12600 real news article dataset obtained from Reuters.com. For classification purpose, the authors have used six machine learning algorithms such as SVM, LR, DT, Stochastic Gradient Descent, kNN and Linear SVM. For this experiment the authors have only consider 2000 articles, 1000 fake and 1000 real. From experimental results it was observed that TFIDF (unigrams) for feature extraction and Linear SVM as a classifier were able to achieve an accuracy of 92%.

Although, the researchers were able to achieve satisfactory results, but as TFIDF is a bag of words model, it does not capture co-occurrence and semantics. To solve this issue of polysemy and synonymy, Blei et al. (2003) recommended LDA as an alternative method for feature extraction technique.

2.3 Feature extraction and topic modelling using Latent Dirichlet Allocation

Many researchers have used LDA as topic modelling and a few researchers have used LDA as feature extraaction.

On Indonesian text Prihatini, Suryawan and Mandia (2018) performed feature extraction using LDA and TFIDF (with Kmeans) individually and compared with each other. The authors explain the reason why TFIDF is applied with K-means is because LDA can cluster the document automatically. The Indonesian text contains 500 news text files which is divided in 5 topics such as News, Automotive, Sport, Technology and Business. For evaluation of the results the authors have used Precision, Recall and F-measure. Initially, TFIDF with K-Means is applied on Indonesian text, the evaluation results were , 0.49 Recall, 11.5 Precision and 0.91 F-measure. Furthermore, LDA is applied on the same Indonesian text, the evaluation results were 0.91 Recall ,0.93 Precision and 0.91 F-measure. From the results, it can be observed that LDA performs better than TFIDF for feature extraction on Indonesian text.

Xue (2019) proposed a technique to handle text retrieval using LDA and Word2vec. In LDA there is global relation of each document to the topics and topics to words whereas Word2vec predicts the target word by taking help of surrounding contextual words. The authors have used 20 Newsgroups dataset which contains 18846 newsgroup documents collected by Ken Lang. The author was able to achieve the best results when the topics were set to 200. To compare the effectiveness of the proposed hybrid method of LDA and Word2vec the authors have used TF-IDF, LDA and Word2Vec individually on the dataset. After applying feature extraction technique, the F1 score of TF-IDF was 0.652, LDA was 0.729, Word2vec was 0.80 and the proposed model's F1 score was 0.814, the hybrid model performed better than the individual feature extraction technique.

Similarly, comparing to work done by Wang, Ma and Zhang (2016), have taken the same 20Newgroup dataset. The authors have used Euclidean distance to calculate the similarity between the topics and the documents. To examine its performance, the authors have used SVM model. The experimental setup was to extract features by feature extraction techniques such as TFIDF, Word2Vec and LDA and then fed to the SVM model. From the results, F1 score of TFIDF + SVM was 0.82, Word2Vec + SVM was 0.717 and LDA + SVM (with 100 topics) was 0.63 and the proposed method by authors LDA + Word2Vec + SVM was 0.803 (with 250 topics). In addition, it worth discussing an interesting fact, the authors highlight that TFIDF performed better than by 2% more accuracy as compare to the authors proposed method. However, TFIDF takes more than 4 times to train with 20000 features as compare to the authors proposed method. Also, TFIDF ignores the semantics between the words. Taking into consideration about the time it takes to run, predicting the accuracy and semantics the authors proposed model performed better as compare to single methods. Comparing the work done by Wang et al. (2016) and Xue (2019), both the authors have used the same hybrid feature extraction model of LDA and Word2Vec. The only difference was Wang employed SVM model to investigate the performance and Xue did not. The f1 score of both the methods were

almost similar, XUE hybrid method F1 score was 0.814 and Wang hybrid F1 score was 0.80.

Kanungsukkasem and Leelanupab (2019) proposed a framework to extract features from news articles using LDA. Financial LDA (FinLDA) is an extension of LDA as it is used for financial time series prediction. FinLDA is a combination of text especially financial time series and news articles using latent Dirichlet allocation to improve prediction of financial time series. The news articles were obtained from Reuters and S&P's 500 index The author extracted features using FinLDA in data preparation phase and to validate the advantages of the features the authors have employed Support Vector Regression and Multi-Layer Perceptron (also denoted as BPNN when back-propagation is used with MLP) machine learning models. The experimental results show that BPNN with FinLDA performed better than SVR. And, the authors also mention that extracting features in data preparation phase helped in making predictions better.

Wang and Xu (2019)proposed a novel approach to improve the performance of the model in detecting fraudulent insurance claims using LDA and deep neural networks. Initially, LDA is used to extract text features from the dataset and then the extracted features are fed to deep neural networks. For comparison purpose the authors have employed traditional machine learning algorithms such as SVM and RF. The authors did the experiment without extracting features with LDA, to check the impact of LDA. The accuracy obtained without LDA was SVM 0.79, RF 0.79 and DNN 0.84. Next, the authors did the experiment with LDA. The accuracy obtained with LDA was SVM 0.78, RF 0.81 and DNN 0.914. It clearly states that employing LDA as a feature extraction technique in data preprocessing stage enhances the performance of the DNN model.

3 Methodology

This section represents methodology employed by this research project. Developed methodology is based on KDD approach (Fayyad et al.; 1996). This research follows the similar methodology employed by Xue (2019) and Ahmed et al. (2017)

3.1 Dataset description

In this research, the dataset used is from real world news articles. The dataset is maintained by author Jruvika and it is obtained from Kaggle.com[2], the author has extracted the dataset from PolitiFact (a Fact checking organization based in USA). The dataset contains 4009 news articles in which 2137 news articles are fake and 1872 news articles real. The label of the real news articles is assigned as 1 and the label of the fake news articles is assigned as 0. The news articles are based on 2016 US presential elections and has following attributes in the dataset:

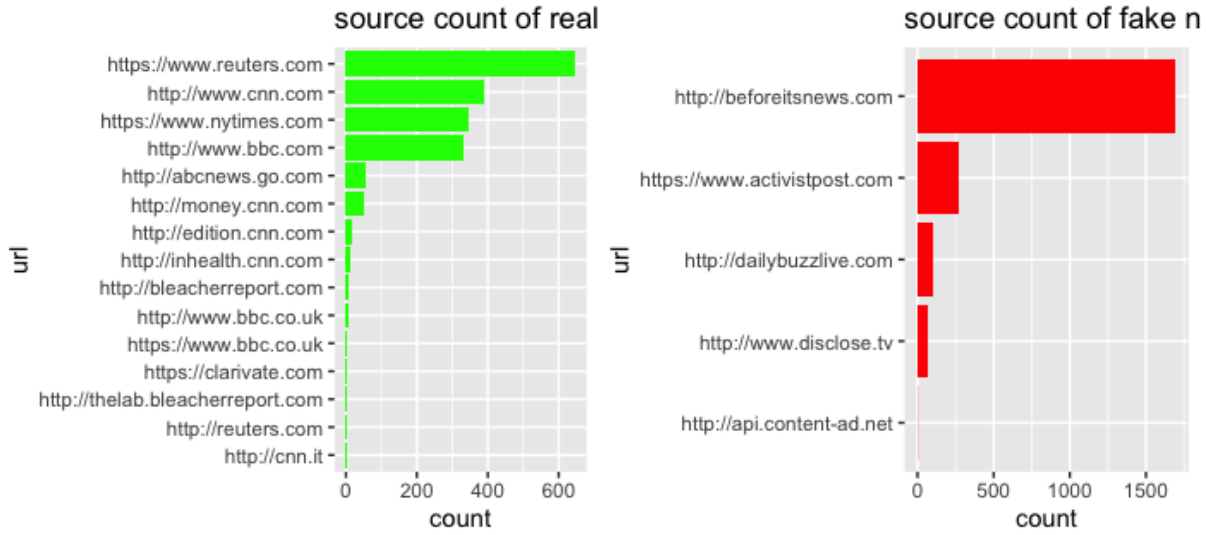
[2]https://www.kaggle.com/jruvika/fake-news-detection

Column name	Description
URL	Urls are the links of original source of news articles
Headline	Headline represents the title of the news-paper,they are minimum facts needed to comprehend the news articles.
Body	Body represents the entire story of the news articles, describing in detail about the place, people involved, etc.
Label	Label represents whether the story is real or fake.

Table 1: Data-set description

3.2 EDA

1. These are main sources where the fake and real news were published



(a) Source of real news

(b) Source of fake news

Figure 2: A figure with two subfigures

2. Welsh t test on the length of the title and text.
   (a) **Title:** Welch test is conducted on the title articles of the news, observed statistically significant difference (p-value = 6.663e-05) between the length of news title of fake and real news. From the test it can be observed the length of the title of fake news is higher than the length of the title of real news. From the figure Figure 3 it can be observed that the center of distribution of length of the title of fake news is centered at mean 7.109967 and the center of distribution of length of the title of real news is centered at mean 6.782585. The Welch t-test shows that the length of real news is shorter than the fake news.

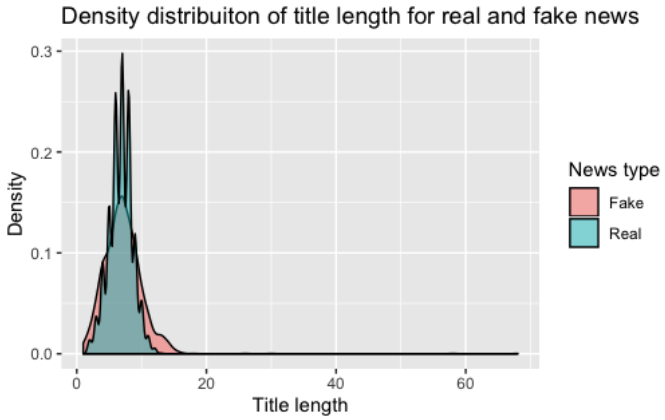


Figure 3: Welsh t test on the length of the title

(b) **Text:** Similarly, Welch test is conducted on the text articles of the news. It was observed there is a statistically significant difference (p-value ¡ 2.2e-16) between the length of text news of fake and real news. From the test it can be observed the length of the title of real news is higher than the length of the title of fake news. From the figure Figure 4 it can be observed that the center of distribution of length of the title of real news is centered at mean 333.9728 and the center of distribution of length of the title of real news is centered at mean 216.1343. The Welch t-test shows that the length of real news is greater than the fake news.

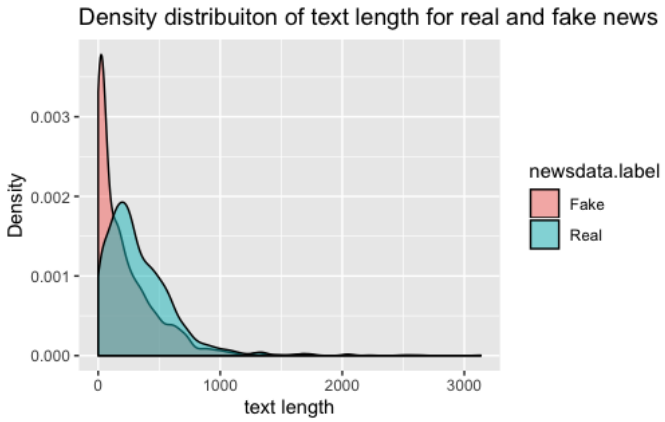


Figure 4: Welsh t test on the length of the text

3.3 Data preprocessing

Pre-processing is a method which translates raw data to meaningful insights. Pre-processing involves steps like removal of stop words, case folding, lemmatization, stemming, tokenization, normalization, noise removal. By removing the unwanted information which have no role in the process will eventually reduce the size of the data. Size of the data plays a major role in decreasing or increasing of computational power, time, cost etc. Also, pre-processing will assist in improving the accuracy of the overall model.

- **Noise removal:** It is one of the most important pre-processing steps that need to be followed for obtaining the best results. The removal of unwanted characters digits and pieces of text, which forms as a disturbance while analysing the text

which is also highly domain dependent in nature. Here, in this step if the text contains source code, header, html formatting, domain specific keyword, etc. will be removed.

- **Removal of stopwords:** Stopwords will generate noise when used as a feature in text classification (Ahmed et al.; 2017). Stopwords words are generally used as a connecting word in English language. Examples of stop word are your, yourself, thus, he, him, his, himself, she, her, hers, herself, it, its, itself etc. Also, articles, prepositions and conjunctions some pronouns are considered as stopwords.
- **Tokenization:** In text classification it is difficult to process sentences. To avoid this issue the data provided is converted into smaller units. This is also called as process of converting human-readable text into machine readable components. In the tokenization process the cases of the word are also taken care. All the words are converted to lower case because in English language normally statement starts with a upper case letter (Prihatini et al.; 2017). It also involves eliminating symbols, punctuation, etc. For example,normal sentence: "Trump wants to build a WALL in Mexico" is converted to "trump wants to build a wall in mexico".
- **Case Folding:** The raw contains a lot of noise, for instance, in place of words there will be numerical values and vice versa, it may also contain special characters and emojis are removed from the dataset. At this phase, the uneven spacing between the words is also removed and upper case are transformed into lowercase (Ahmed et al.; 2017). For instance, "Amer!c@" is converted to "america".
- **Stemming:** The following step after removing the stopwords is to do stemming. Stemming converts words to its word root. For example, the words "farming", "farm" and "farmer" will be reduced to the word "farm."

3.4 Proposed hybrid method using LDA and TFIDF

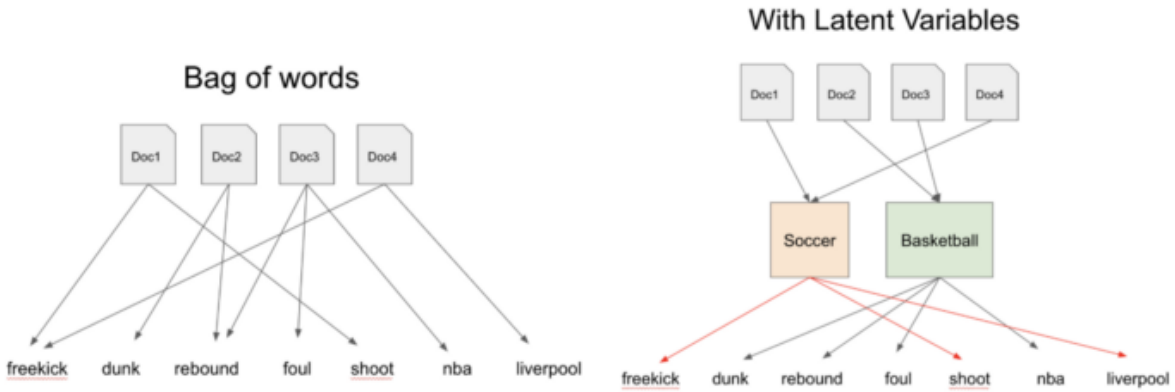


Figure 5: Figure on the right represents LDA approach and figure on left represents traditional bag of words approach

By using traditional bag of words approach for extracting features from the text, each document is mapped to the words token with the help of a document term matrix. This

results in a sparse matrix with many zeros. Many parameters are needed to be evaluated when considering such a matrix as a input to the machine learning models, this often results in noisy information.

Whereas in LDA approach, instead of mapping each text document to word token directly, the latent topics are introduced as bridges. Within a corpus, each document is identified by Dirichlet Distribution over the latent topics and each topic is identified by another Dirichlet Distribution over all the word tokens.

Futhermore, LDA does not consider the importance of a word to a document in a collection or corpus. Thus, in this work using TFIDF term weighting schemes are introduced to provide weight to the words.

In this research, first, weights are assigned for the terms using Term Frequency- Inverse Document Frequency. Using top weights of top 40, 50 terms, this acts as a input to LDA. Futher, LDA is applied on the extracted features by TFIDF. Later, these extracted features from TFIDF and LDA will be analysed using different supervised classification models.

3.5 Classification of models

In this research, we have modelled the fake news detection problem as binary classification problem. F : $\in \rightarrow \{0,1\}$

$$F(x) = \begin{cases} 1, & \text{if } x \; is \; a \; fake \; news \; article \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In this research, we are using Linear SVM, Radial SVM, Naive Bayes and Random Forest. Researchers Agudelo et al. (2018),Wang (2017) from the literature review have employed SVM to detect fake news. Similarly, Granik and Mesyura (2017) have employed Naive Bayes, have employed Random to detect fake news.

3.6 Metrics

Many researchers have considered fake news as a classification problem.This project follows Shu et al. (2017) procedure for evaluating metrics. In this research, confusion matrix is used to evaluate the performance of the classification model. Developed binary fake news classifier will produce results in four categories:

1. **True Positive:** When predicted fake news articles are actually marked as fake news.
2. **True Negative:** When predicted true news articles are actually marked as true news.
3. **False Negative:** When predicted true news articles are actually marked as fake news.
4. **False Positive:** When predicted fake news articles are actually marked as true news.

For evaluation of performance of the models accuracy, precision, recall and F1 score is taken into consideration.

Precision: Precision tells us about the success probability of producing a correct positive class classification Precision estimates the fraction all the identified fake news that are marked as fake news. It is calculated as a ratio of true positives to total number

of positive values. As the data is little skewed, there are chances that by making fewer positive predictions we can get high precision.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Figure 6: Equation to calculate Precision

Recall: Therefore, in this scenario Recall can be employed. It is used to measure the sensitivity. It is calculated as news articles that are marked fake new over news articles that are predicted to be fake.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Figure 7: Equation to calculate Recall

F1: F1 is used to measure the total performance of the prediction for detecting fake news. It is the harmonic mean of Precision and Recall.

$$f_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Figure 8: Equation to calculate F1

Accuracy: Accuracy is the ratio of correctly predicted observation to the total observations

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Figure 9: Equation to calculate Accuracy

4 Design Specification

4.1 Process flow diagram

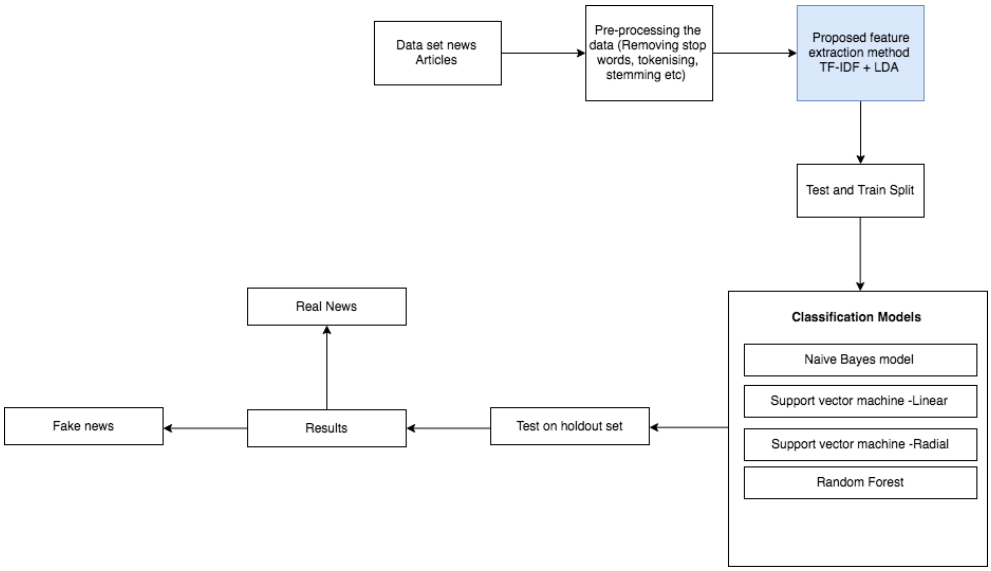


Figure 10: Process flow diagram

4.2 TFIDF

TFIDF : As mentioned in the introduction, TFIDF assigns weight according to importance of the terms. Rarer the term, higher is the score of the term in the document set. For instances, if TFIDF assigns weight to a word (t) in a document (d) that is given by,

$$tf\text{-}idf_{t,d} = tf_{t,d} * idf_t$$

Figure 11: TF-IDF equation

1. The weight is highest when the term t appears many times is small set of documents.
2. The weight is lower when the word appears many few times in many documents or in a document.
3. The weight is lowest when it appears in all the documents.

4.3 LDA model

Figure 1 represents statistical model. LDA models works in three stages. Firstly, the model assumes that topics (K) is been initialised beforehand. Secondly, the algorithm goes through each document and assigns the words to a temporary topic in random way as in the next phase it will updated. Thirdly, the algorithm will run in loop through each word in the document and will update words in topics for better results.

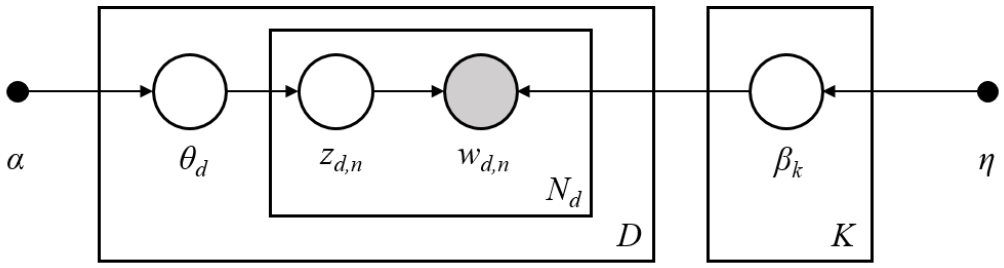


Figure 12: LDA model Adapted from Blei et al. (2003)

α – It is the distribution of topics per document
β – distribution of words per topics
θ – Topic distribution for topic K
K- Number of topics
D – D represents the corpus
Z – per word topic assignment
W – observed word
N – Number of words per document

5 Implementation

This chapter represents how the research project was implemented and classified using machine learning algorithms to detect fake news. Also, it discusses features. The research project is done in R Studio Cloud where the R version is 3.6.0, platform is x86_64-pc-linux-gnu (64-bit) and it is running under Ubuntu 16.04.6 LTS. Data is collected from Kaggle.com and downloaded all the packages required using install.packages() function. Installed libraries such as tm library is text mining framework, tidytext for text manipulation, tidyverse for preprocessing and visualisation.SnowballC library is used for stemming, Lda function from topic model library, ggplot for graphs and visualization First, read the data by importing using read.csv() function. The dataset contains 4009 news articles in which 2137 news articles are fake and 1872 news articles real. The label of the real news articles is assigned as 1 and the label of the fake news articles is assigned as 0.The raw text needs to be converted into a format that can be used for extracting information.Imported the dataset in R using read.csv function. Generated ID column using seq.int(nrow(data)) function and removed the unwanted extension from URL after .com, .uk, .tv, .net. Then for simplicity changed the names of columns where Headline was changed to title and body was changed to text. Normalization helps to reduce the size of the data before analysing the text. Using tm package created a corpus function called corpus_all() to remove special characters, numbers, transform words to lower case, remove punctuation's, remove numbers, remove extra space, and stemming. Further, to make sure everything cleaning of unwanted information created a custom clean function clean_text() as explained in section 3.3. Later,to analyse the text files, all the files needs to be converted into Document Term Matrix(DTM). Each term is depicted in a given document as columns and rows in a DTM depicts the documents. For feature extraction using TFIDF, top 50 term which are having higher weight is considered and for feature extraction using LDA topics(k=2, k=4) is considered. As we increases the number of topic the results declines. This is due to the taking just top 50 features from TFIDF.For modelling, Caret package was used. kLAR package for Naïve Bayes – under caret, for

SVM radial and linear kernel is used, also, employed random forest under caret package. XGBoost package – Expand grid search to tune.

Before training the machine learning algorithms, data is splitted in 80%-20% by data partition function. To avoid overfitting 5 fold CV is used. Lastly, using predict function, accuracy of machine learning is calculated. Confuison matrix is used to evaluate the performance of classification model.

6 Evaluation

6.1 Baseline model Naive Bayes

Accuracy	**Precision**	**Recall**	**F1**
47.54	49.7	20.20	38.90

Table 2: Naive Bayes

6.2 Evaluation and Results of Linear SVM

The score function (which computes score for a new variable) in SVM-Linear model is linear and parametric. These functions are initially introduced to model for satisfying the concept of margin maximization. Further, it works well with high dimensional data and this model gives best result if the data can be linearly separable. As the kernels are not introduced here, it takes less computation power and time than usual. Also, Ahmed et al. (2017) were able to get best results with Linear SVM in detecting fake news.

Feature extraction technique	**Accuracy**	**Precision**	**Recall**	**F1**
TFIDF	66.24	69.84	68.2	68.83
LDA (k=2)	49.54	48.43	38.17	49.97
LDA (k=4)	50.1	51.28	35.23	48.32
TFIDF + LDA	51.91	52.35	41.33	46.89

Table 3: Linear SVM

TFIDF was able to achieve highest accuracy of 66.24%, Precision of 69.84%, Recall of 68.3% and F1 of 68.8%. On the other hand, lowest accuracy with LDA(k=4) topic with 50.1%.

6.3 Evaluation and Results of Radial SVM

The SVM-Radial model uses radial kernel to classify the data which is not linearly separable. With the introduction of kernels, the model will be flexible enough to choose the threshold in classifying the data.Further, If the parameters are chosen appropriately then the SVM model with kernels will perform a good out-of-sample generalization. Also, the advantage of SVM over neural networks is that, it provides a solution which is unique, and this is because the optimally problem is curved in nature which is faced most of the times while analyzing the text.

Feature extraction technique	Accuracy	Precision	Recall	F1
TFIDF	59.69	83.02	32.22	47
LDA (k=2)	49.9	49.85	49	42.21
LDA (k=4)	49.56	51.44	36.2	43.42
TFIDF + LDA	54.23	53.11	45.1	57.33

Table 4: Radial SVM

TFIDF was able to achieve highest accuracy of 59.69% and Precision of 83%, LDA(k=2) was able to achieve highest Recall of 49% and TFIDF + LDA highest F1 of 57.33%.

6.4 Evaluation and Results of Naive Bayes

Naive Bayes model was build by using bags of words model (TF).This machine learning model can work better even with a less amount of training data. Also, it is not sensitive to features which are irrelevant. Its highly scable, and it scales linearly as the number of predictors increases.

Feature extraction technique	Accuracy	Precision	Recall	F1
TFIDF	72.04	72.64	78.34	75.38
LDA (k=2)	52.34	53.6	43.22	38.43
LDA (k=4)	53.22	52.55	43.59	47.87
TFIDF + LDA	55.65	54.33	46.12	54.78

Table 5: Naive Bayes

TFIDF was able to achieve highest accuracy of 78.34%, Precision of 72.64%, Recall of 78.3% and F1 of 75.38%.

6.5 Evaluation and Results of Random Forest

The primary idea behind using random forest model in machine learning is that,the process of combining decisions of many decision trees helps to overcome the issue of over-fitting. Each individual tress gives its own output by considering random subset of features of the data. Even with missing values present in the data they are able to maintain the accuracy.

Feature extraction technique	Accuracy	Precision	Recall	F1
TFIDF	76.07	82.1	71.2	76.1
LDA (k=2)	51.23	49.2	43.61	45.98
LDA (k=4)	53.63	54.47	41.16	48
TFIDF + LDA	56.21	59.89	44.4	51.12

Table 6: Random Forest

TFIDF was able to achieve highest accuracy of 76.07%, Precision of 82.1%, Recall of 71% and F1 of 76.1%.

7 Discussion

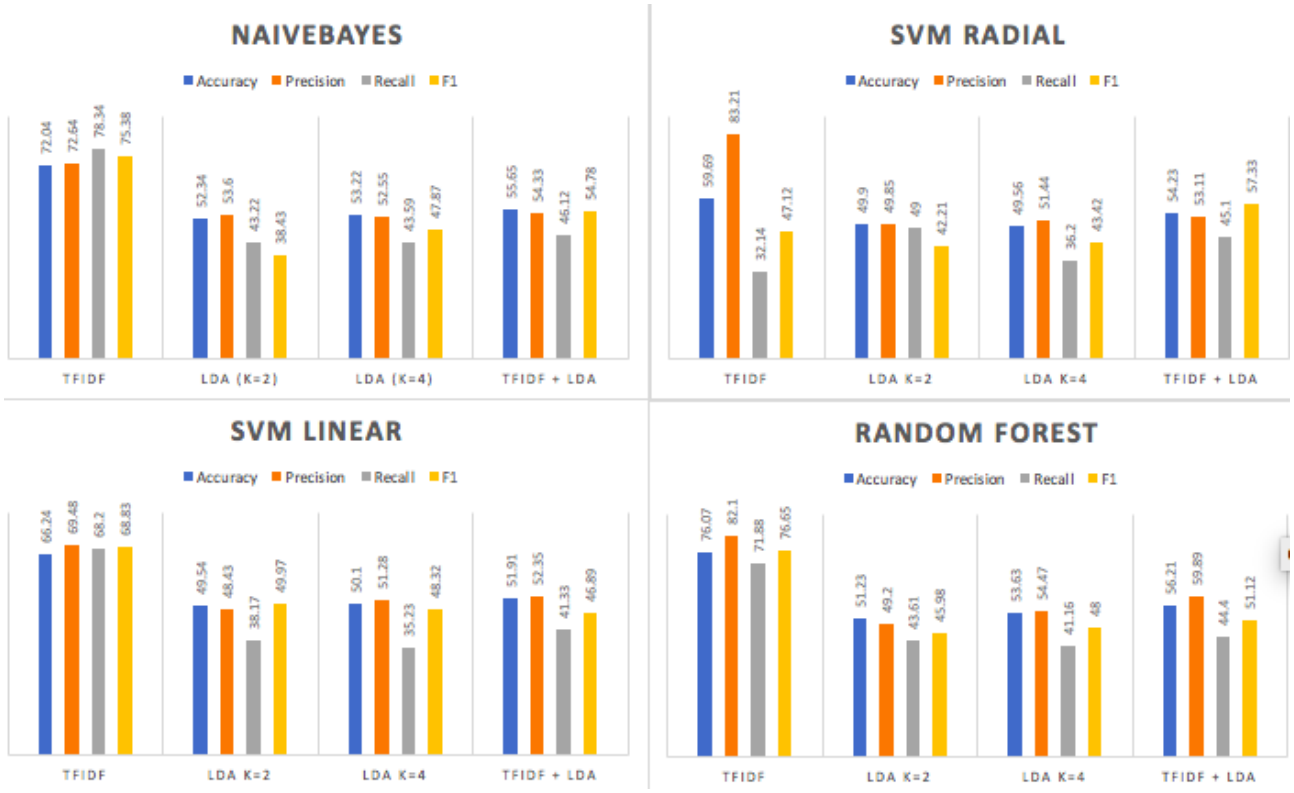


Figure 13: Discussion

It can be observed that Fig 13 with Naive Bayes, TFIDF was able to achieve highest accuracy of 76.07%, Precision of 82.1%, Recall of 71% and F1 of 76.1%. And, even with Linear SVM, TFIDF was able to achieve highest accuracy of 59.69% and Precision of 83%,whereas in Linear SVM LDA(k=2) was able to achieve highest Recall of 49% and TFIDF + LDA highest F1 of 57.33%. Futhermore, TFIDF with Radial SVM was able to achieve highest accuracy of 66.24%, Precision of 69.84%, Recall of 68.3% and F1 of 68.8%. On the other hand in Radial SVM, lowest accuracy with LDA(k=4) topic with 50.1%. Lastly, TFIDF with Random FOrest was able to achieve highest accuracy of 78.34%, Precision of 72.64%, Recall of 78.3% and F1 of 75.38%. Additionally, it can observed that highest precision was achieved by TFIDF with SVM Radial with 83.21 precision. And, lowest precision was achieved by Linear SVM with LDA(k=2) of 48.43. Also, Additionally, it can observed that highest accuracy was achieved by TFIDF with Random Forest with 76.07 accuracy. And, lowest accuracy was achieved by Linear SVM with LDA(k=2) of 49.9.

From the above discussion, it can be observed although LDA performed better than baseline classifier. TFIDF outperformed LDA and out proposed methodology. LDA did not perform as good as TFIDF one of the reason is that, in this project we had considered top 50 features extracted from TFIDF. It can be concluded that, from the experiment we saw LDA does not work well with smaller text.

8 Conclusion and Future Work

In this research, we proposed a hybrid feature extraction technique using TF-IDF and LDA. In our proposed hybrid method, first, weights are assigned for the terms using Term Frequency- Inverse Document Frequency. Then top 50 terms are selected based on their weights, this acts as a input to LDA. Futher, LDA is applied on the extracted features by TFIDF. Later, these extracted features from TFIDF and LDA will be analysed using different supervised classification models. Also, for comparison purpose we extracted features individually by TFIDF and LDA along with machine learning algorithms. For this research project we have employed Linear SVM, Radial SVM, Random Forest and Naive Bayes. Experimental results shows that although our proposed method was able to perform better than baseline classifier. But, Random Forest (76.07% accuracy) works the best with the features extracted by TF-IDF model individually in the process of fake news classification. LDA didnot perform as good as TFIDF one of the reason is that, in this project we had consideredtop 50 features extracted from TFIDF. It can be concluded that, from the experimentwe saw LDA does not work well with smaller text In future work, we would create an ensemble approach by combining title and text together and taking a dataset where the length of articles are long.

Acknowledgement

I would like to thank my mentor Ms. Sidra Bashir for her guidance. She always encouraged me to steer in right direction and shared her feedback throughout the research project. Also, I would like to take this opportunity to thank all staff members of Data Analytics.

References

Agudelo, G. E. R., Parra, O. J. S. and Velandia, J. B. (2018). Raising a model for fake news detection using machine learning in python, *in* S. A. Al-Sharhan, A. C. Simintiras, Y. K. Dwivedi, M. Janssen, M. Mäntymäki, L. Tahat, I. Moughrabi, T. M. Ali and N. P. Rana (eds), *Challenges and Opportunities in the Digital Era*, Springer International Publishing, Cham, pp. 596–604.

Ahmed, H., Traore, I. and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques, *in* I. Traore, I. Woungang and A. Awad (eds), *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, Springer International Publishing, Cham, pp. 127–138.

Ahuja, R., Chug, A., Kohli, S., Gupta, S. and Ahuja, P. (2019). The Impact of Features Extraction on the Sentiment Analysis, *Procedia Computer Science* **152**: 341–348.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050919306593*

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *The Journal of Machine Learning Research* **3**: 993–1022.
URL: *http://dl.acm.org/citation.cfm?id=944919.944937*

Bourgonje, P., Moreno Schneider, J. and Rehm, G. (2017). From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles,

Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, Association for Computational Linguistics, Copenhagen, Denmark, pp. 84–89.
URL: *https://www.aclweb.org/anthology/W17-4215*

Chiu, J., Gokcen, A., Wang, W. and Yan, X. (2013). Classification of fake and real articles based on support vector machines language and statistics spring 2013.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data, *Commun. ACM* **39**(11): 27–34.
URL: *http://doi.acm.org/10.1145/240455.240464*

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, pp. 1163–1168.
URL: *https://www.aclweb.org/anthology/N16-1138*

Fontanarava, J., Pasi, G. and Viviani, M. (2017). Feature analysis for fake review detection through supervised classification, *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 658–666.

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

Granik, M. and Mesyura, V. (2017). Fake news detection using naive bayes classifier, *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903.

Hardalov, M., Koychev, I. and Nakov, P. (2016). In Search of Credible News, *in* C. Dichev and G. Agre (eds), *Artificial Intelligence: Methodology, Systems, and Applications*, Springer International Publishing, Cham, pp. 172–180.

Jain, A. and Kasbe, A. (2018). Fake news detection, *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–5.

Kanungsukkasem, N. and Leelanupab, T. (2019). Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction, *IEEE Access* **7**: 71645–71664.
URL: *https://ieeexplore.ieee.org/document/8726415/*

Kotteti, C. M. M., Dong, X., Li, N. and Qian, L. (2018). Fake news detection enhancement with data imputation, *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 187–192.

Mukherjee, A., Venkataraman, V., Liu, B. and Glance, N. (2013). Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews.
URL: *https://pdfs.semanticscholar.org/4c52/1025566e6afceb9adcf27105cd33e4022fb6.pdf*

Prihatini, P. M., Putra, I. K. G. D., Giriantari, I. A. D. and Sudarma, M. (2017). Fuzzy-Gibbs latent Dirichlet allocation model for feature extraction on Indonesian documents, *Contemporary Engineering Sciences* **10**: 403–421.
URL: *http://www.m-hikari.com/ces/ces2017/ces9-12-2017/7325.html*

Prihatini, P. M., Suryawan, I. K. and Mandia, I. (2018). Feature extraction for document text using Latent Dirichlet Allocation, *Journal of Physics Conference Series*, Vol. 953 of *Journal of Physics Conference Series*, p. 012047.

Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H. (2017). Fake news detection on social media: A data mining perspective, *CoRR* **abs/1708.01967**.
URL: *http://arxiv.org/abs/1708.01967*

Srivastava, A., Rehm, G. and Moreno Schneider, J. (2017). DFKI-DKT at SemEval-2017 Task 8: Rumour Detection and Classification using Cascading Heuristics, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, pp. 486–490.
URL: *https://www.aclweb.org/anthology/S17-2085*

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* .
URL: *http://dx.doi.org/10.18653/v1/P17-2067*

Wang, Y. and Xu, W. (2019). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, *Decision Support Systems* **105**: 87–95.
URL: *http://www.sciencedirect.com/science/article/pii/S0167923617302130*

Wang, Z., Ma, L. and Zhang, Y. (2016). A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2vec, *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, IEEE, Changsha, China, pp. 98–103.
URL: *http://ieeexplore.ieee.org/document/7866114/*

Xu, K., Wang, F., Wang, H. and Yang, B. (2018). A first step towards combating fake news over online social media, *in* S. Chellappan, W. Cheng and W. Li (eds), *Wireless Algorithms, Systems, and Applications*, Springer International Publishing, Cham, pp. 521–531.

Xue, M. (2019). A Text Retrieval Algorithm Based on the Hybrid LDA and Word2vec Model, *2019 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, pp. 373–376.

Zhang, X. and Ghorbani, A. A. (2019). An overview of online fake news: Characterization, detection, and discussion, *Information Processing & Management* .
URL: *http://www.sciencedirect.com/science/article/pii/S0306457318306794*