Research Methods

Project Report



Identification of Fake News on Social Media based on linguistic leakages

Submitted by: Submitted to:

Syed Waseem Hussain Prof. Pranav Ranjan

Date:

14/06/2018

**Abstract**

With the massive explosion of internet connectivity and the advent of social media in India, large volumes of fake news, are produced online for a variety of purposes, such as personal, financial and political gain. This extensive spread of fake news through social media has the potential for extreme negative impacts on the society. Therefore detecting “fake news” is of prime concern in order to make sure that the news that reaches the people through social media and other modes are valid and reliable. The biggest problem that is that it is impossible to manually verify or filter the news content at such humongous sizes.

There are conceptual level of research available where distinction has been made between three types of fake news: serious fabrications, hoaxes and satire. The proposed research is based on the fact that the liars or hoax spreaders around the world use their language strategically to avoid being caught. Despite their control, language ‘leakage’ occurs with certain verbal aspects. Identifying such patterns of leakage using advanced predictive analysis and Natural Language Processing (NLP) techniques in order to automate this work and make it easier and efficient is a core part of the research.

The data of the proposed research will be divided into two parts, one being the genuine news articles from verified news portals, websites and blogs. The other part of the data will consist of fake news articles circulated on social media belonging to business, entertainment and politics domains.

The methodology used in the research will be Pre-processing the text using Python packages, to analyze and model text by converting text to features using techniques such as TF-IDF or Word2Vec, utilizing Naïve Bayes, Logistic Regression and Random Forest for the implementation part. Cross-validation and a grid search to find the best parameters for the TF-IDF algorithm and each individual model will be used. Classification model giving the highest accuracy among these 3 will be selected to detect fake news on social media platforms for further research.

Key Words: Social Media, Fake News, Society, Predictive Modelling, Natural Language Processing.

# Project Introduction

Most of our times are spent interacting online through social media platforms, more and more individuals tend to look for news on social media rather than traditional news groups. The motives for this transformation in consumption activities are due to:

* It’s repeatedly timelier and less costly to feed news on social media compared with traditional news media, such as newspapers or television.
* It’s easier to further share, comment, and discuss the news with friends or other readers on social media.

According a research report (Shu, Sliva, Wang, Tang, & Liu, 2017), It is found that 62% of U.S. adults get news on social media in 2016, while in 2012, only 49% reported seeing news on social media and it is also known that social media has now outperforms television as the major news source. Despite the advantages provided by social media, the quality of news on social media is lower than traditional news groups. However, because it is less expensive to provide news online and much faster and easier to circulate through social media, large volumes of fake news, i.e., those news articles with intentionally false information, are produced online for a variety of purposes, such as financial and political gain. It was estimated that over 1 million tweets are related to fake news.

The biggest weapon of the present is the Mass Media. It has the power to influence the people. The scenario is in such a way that people are blindly believing whatever news that the social media or the other modes of mass communication present to them. There are a lot of organizations and individuals who are ready to spread “fake news” through this mass media for their personal gains.

According to a report (Murali Krishnan, 2017), in India the problem is twice as compounded with the massive explosion of internet connectivity. Studies show that there are now 355 million internet users in India, roughly 27% of the population. India now has the second highest number of internet users in the world after China. A study by the Internet and Mobile Association of India and IMRB International, a market research firm, concluded that 77% of urban internet users and 92% of rural users consider mobile phone as the primary device for accessing the internet, largely driven by availability and affordability of smartphones.

Some telephone companies offer free data to users with a data cap of 4 GB per day. This has forced other network providers to come out with competitive plans offering data at much cheaper rates. What we then see is an increasing trend of circulation of fake news and photographs. India is one of the biggest markets (Surabhi Agarwal, 2018) for several social media and communication companies it has 160 million of WhatsApp's one billion-plus monthly active users, 148 million Facebook users and over 22 million Twitter accounts. There are two major factors driving this fake news surge. First is the trend of declining smartphone prices over the past couple of years. And, second is the fall in internet data prices.

This extensive spread of fake news has the potential for extreme negative impacts on the society. Therefore detecting “fake news” is of prime concern in order to make sure that the news that reaches the people through social media and other modes are valid and reliable. The biggest problem that is that it is impossible to manually verify or filter the news content at such humongous sizes. There are conceptual level of research (Rubin, Conroy, Chen, & Cornwell, 2016) available where distinction has been made between three types of fake news: serious fabrications (i.e. news items about false and no existing events or information such as celebrity gossip), hoaxes (i.e. providing false information via, for example, social media with the intention to be picked up by traditional news websites) and satire (i.e. humorous news items that mimic genuine news but contain irony and absurdity).Therefore, one needs to resort to advanced predictive analysis and Natural Language Processing (NLP) techniques in order to automate this work and make it easier and efficient. Our research is based on the fact that the liars or hoax spreaders around the world use their language strategically to avoid being caught. Despite their control, language ‘leakage’ occurs with certain verbal aspects. Identifying such patterns of leakage is a core part of the research.

**Literature Review**

(Rosas, et al., 2018) introduced two novel datasets for the task of fake news detection, covering seven different news domains. They introduced two new fake news datasets, one obtained through crowdsourcing and covering six news domains, and another one obtained from the web covering celebrities. They developed classification models that rely on a combination of lexical, syntactic, and semantic information, as well features representing text readability properties. Their best performing models achieved accuracies that are comparable to human ability to spot fake content. They used a linear SVM-classifier and five-fold cross-validation. Results showed that language used when reporting fake content uses more social and positive words, expresses more certainty and focuses on present and future actions. Moreover, the authors of fake news use more adverbs, verbs, and punctuation characters than the authors of legitimate news.

(Rubin, 2016) concluded that social media requires content verification analysis with a combination of previously known approaches for deception detection, as well as novel techniques for debunking rumors, credibility assessment, factivity analysis and opinion mining. Hybrid approaches may include text analytics with machine learning for deception detection, network analysis for rumor debunking and should incorporate world knowledge databases to fully take advantage of the linguistic, interpersonal, and contextual awareness.

(Bajaj, 2017) used NLP to build a classifier that can predict whether a piece of news is fake based on its content. Several different models were implemented. Logistic Regression, Two-Layer Feedforward Neural Network, Recurrent Neural Network, Long Short-Term memories, Gated Recurrent Units, Bidirectional RNN with LSTMs, Convolutional Neural Network with Max Pooling and Attention-Augmented Convolutional Neural Network are the models used in the classifier. Only the feedforward network and the RNNs with complex activation units achieved a recall greater than 0.7, indicating that the models were able to extract some subset of relevant features to classify more of the positive examples correctly.

(Shu, et al., 2018) collected verified fake news from news portals and tweets using twitter advanced search APIs. They propose Social Article Fusion (SAF) model that uses the linguistic features of news content and features of social context to classify fake news. Auto-encoders are widely used to represent the text content in lower dimensional space. So, they also used auto encoder to capture the text content of the news articles in the lower dimensional space z. news. The interactions of the users on the social media change over the period of time. To capture the temporal engagements of the users with the fake news they have used Recurrent Neural Networks (RNN). The output of the RNN is considered as the social context feature for the classification. Long Short-Term Memory (LSTM) has been used as it solves the long-range dependencies and vanishing gradient problem. To test the quality of the learned features they tried the classification task using the learned feature with standard machine learning algorithms like Support Vector Machine, Logistic Regression and Naïve Bayes.

(Rubin, et al., 2017) discussed three types of fake news. Serious fabrications (uncovered in mainstream or participant media, yellow press or tabloids); large-scale hoaxes and humorous fakes (news satire, parody, game shows). They concluded that Serious fabricated news may take substantial efforts to collect, case by case. Journalistic fraudsters may face harsh consequences for dishonest reporting, and are likely to exhibit cues of deception akin to “verbal leakages” in other contexts (such as law enforcement or CMC). Large-scale hoaxing attacks are creative, unique, and often multi-platform, which may require methods beyond text analytics (e.g., network analysis). Humorous news provides a steady stream of data, but their writers’ intentions to entertain, mock, and be absurd may interfere with binary text classification techniques, especially if algorithms pick up cues of believability, sensationalism, or humor instead of cues for deception. As an important development at an intersection of LIS, NLP, big data and journalism, fake news detection (of the three identified types) holds promise in automated news verification and online content credibility assessment.

(Rubin & Conroy, 2012) in their research to classify deceptive news took a sample of 90 computer mediated personal stories with varying levels of deception. Each story had 10 associated human deception level judgments, confidence scores, and explanations. In total, 990 unique respondents participated in the study. Three approaches were taken to the data analysis of the sample: human judges, linguistic detection cues, and machine learning. Binary Text classification was used as a Machine learning algorithm. Comparable to previous research results, human judgments achieved 50–63 percent success rates, depending on what was considered deceptive.

(Ali & Levine, 2008) examined differences in language usage as a function of message veracity and speech act type. A quasi-experiment crossed truthful and deceptive messages with confessions and denials in an induced cheating situation. Transcribed messages were analyzed with Linguistic Inquiry and Word Count (LIWC) software. The data was analyzed with a series 2 x 2 independent groups ANOVAs with one ANOVA for each of the 34 coded linguistic features. Relative to honest participants, liars exhibited significantly fewer negative emotions, less discrepancy, fewer modal verbs, more modifiers, and they spoke longer. Relative to confessions, denials were characterized by shorter sentences, more negations, greater discrepancy, fewer past tense verbs, and more present tense verbs. No statistically significant interactions were observed. The effect sizes ranged from 13% to 27% of the variance, and thus the statistically significant effects reflected substantial differences.

(Shu, et al., 2018) provided a fake news data repository FakeNewsNet, which contained two comprehensive datasets that includeed news content, social context, and dynamic information. They presented a comprehensive description of datasets collection, demonstrated an exploratory analysis of this dataset from different perspectives, and discussed the benefits of FakeNewsNet for potential applications on fake news study on social media.

(Bachenko & Fitzpatrick, 2008) used natural language processing to identify deceptive and non-deceptive passages in transcribed narratives by motivating an analysis of language-based deception that relies on specific linguistic indicators to discover deceptive statements. The indicator tags were assigned to a document using a mix of automated and manual methods. An interpreter automatically discriminated between deceptive and truthful statements based on tag densities. The texts used in the study came entirely from “real world” sources—criminal statements, police interrogations and legal testimony. Classification and Regression Tree techniques suggested that the approach was feasible, with the model able to identify 74.9% of the T/F propositions correctly. Implementation of an automatic tagger with a large subset of tags performed well on test data, producing an average score of 68.6% recall and 85.3% precision when compared to the performance of human taggers on the same subset. The results strongly suggested that linguistic cues provide a guide to deceptive areas of a text.

(Agren & Agren, 2018) automated the challenging task of stance detection in the news domain, specifically determining the relative stance of an article towards a claim stated in an associated headline, making use of the labelled dataset delivered for supervision in the first stage of the Fake News Challenge. The FNC-1 dataset consisted of a training set and a test set in which there are 49,972 headline-article pairs in the training set and an additional 25,413 pairs in the test set. Models which were used were Parallel encoder, Conditional encoder and Neural Attention. The results reported that ensemble classifier performed the best on the FNC-1 test set. The results reported in this thesis indicated that the RNN based classifiers are weak on the subtask of distinguishing if a headline-article pair is related or not. They found that the RNN based models suffer from overfitting. In addition, much of the information in an article is considered noise and does not contribute with any useful content for making a certain classification.

(Wang, 2017) investigated automatic fake news detection based on surface-level linguistic patterns. They designed a novel, hybrid convolutional neural network to integrate metadata with text. They showed that this hybrid approach can improve a text-only deep learning model. They used a publicly available data set called LIAR. The LIAR dataset included 12.8K human labeled short statements from POLITIFACT.COM’s API, and each statement was evaluated by a POLITIFACT.COM editor for its truthfulness. They used five baselines: a majority baseline, a regularized logistic regression classifier (LR), a support vector machine classifier (SVM), a bi-directional long short-term memory networks model (Bi-LSTMs) and a convolutional neural network model (CNNs). For LR and SVM, they used the LIBSHORTTEXT toolkit6 , which was shown to provide very strong performances on short text classification problems. For Bi-LSTMs and CNNs, they used TensorFlow for the implementation. They used pre-trained 300 dimensional word2vec embeddings from Google News to warm-start the text embeddings. Results showed that when combining meta-data with text, significant improvements can be achieved for fine-grained fake news detection.

(Vedova, et al., 2018) proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing their already high accuracy by up to 4.8%. They implemented their method within a Facebook Messenger chatbot and validated it with a real-world application, obtaining a fake news detection accuracy of 81.7%. The dataset consisted of the public posts and posts’ likes of a list of Facebook pages. The final dataset was composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2,300,00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes and 6,577 (42.4%) are non-hoaxes. The second and third datasets came from the FakeNewsNet dataset provided by PolitiFact and BuzzFeed. Three methods were used on the datasets: Content-based (CB), Logistic regression (LR) on social signals and Harmonic boolean label crowdsourcing (HC) on social signals. They then tested the chatbot with a completely independent set of news, whose content was not part of the three previously mentioned datasets. The results confirmed their hypothesis: although social-based methods offer good performances, these performances get worse for news items that collect only few social interactions, and therefore the combination with content-based approaches can increase the overall detection accuracy.

(Shu, et al., 2018) reviewed network properties for studying fake news, introduced popular network types and proposed how these networks can be used to detect and mitigate fake news on social media. According to them the news dissemination ecosystem on social media involves three dimensions, a content dimension, a social dimension, and a temporal dimension. They built homogeneous and heterogeneous networks within a specific dimension and across dimensions. The dimensions of fake news dissemination revealed mutual relations and dependencies that can form different types of networks.

(Gupta, et al., 2018) represented a news article as a 3-mode tensor of the structure - and proposed a tensor factorization based method to encode the news article in a latent embedding space preserving the community structure by modeling the echo-chambers as closely-connected communities within the social network. They also proposed an extension of the above method, which jointly models the community and content information of the news article through a coupled matrix-tensor factorization framework. They empirically demonstrated the efficacy of their method for the task of Fake News Detection over two real-world datasets. To benchmark their method, they considered a number of baselines, those can be broadly categorized into two types: Content-Based Baselines and Content + Social Engagement Based Baselines. Further, they validated the generalization of the resulting embeddings over two other auxiliary tasks, namely: News Cohort Analysis and Collaborative News Recommendation. Their proposed method outperforms appropriate baselines for both the tasks, establishing its generalization.

(Tacchini, et al., 2017) showed that Facebook posts can be classified with high accuracy as hoaxes or non-hoaxes on the basis of the users who “liked” them. They presented two classification techniques, one based on logistic regression, the other on a novel adaptation of Boolean crowdsourcing algorithms. On a dataset consisting of 15,500 Facebook posts and 909,236 users, they obtained classification accuracies exceeding 99% even when the training set contained less than 1% of the posts. They further showed that their techniques are robust: they worked even when they restricted their attention to the users who like both hoax and non-hoax posts. These results suggested that mapping the diffusion pattern of information can be a useful component of automatic hoax detection systems.

(Ahmed, et al., 2017) proposed a fake news detection model that uses n-gram analysis and machine learning techniques. They investigated and compared two different features extraction techniques and six different machine classification techniques. They collected news articles from Reuters.com for real news articles. As for the fake news, they were collected from a fake news dataset on kaggle.com. SVM, LSVSM, KNN, DT, SGD and LR models were used. Results showed that Linear-based classifiers (Linear SVM, SDG, and Logistic regression) achieved better results than nonlinear ones. The highest accuracy was achieved using Linear SVM as 92%.

(Shu, et al., 2017) presented 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. They also discussed related research areas, open problems, and future research directions for fake news detection on social media. News content were mostly linguistic-based and visual based. There were three major aspects of the social media context that they represented: users, generated posts, and networks. Their dataset included news from BuzzFeedNews, LIAR, BS Detector and • CREDBANK. They explored the fake news problem by reviewing existing literature in two phases: characterization and detection. In the characterization phase, they introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, they reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction.

(Ruchansky, et al., 2017) in their work stated that there are three generally agreed upon characteristics of fake news: the text of an article, the user response it receives, and the source users promoting it. They proposed a model that combined all three characteristics for a more accurate and automated prediction. Specifically, they incorporated the behavior of both parties, users and articles, and the group behavior of users who propagate fake news. They proposed a model called CSI which is composed of three modules: Capture, Score, and Integrate. The module is based on the response and text; it uses a Recurrent Neural Network to capture the temporal pattern of user activity on a given article. The second module learns the source characteristic based on the behavior of users, and the two are integrated with the third module to classify an article as fake or not. Experimental analysis on real-world data demonstrates that CSI achieves higher accuracy than existing models, and extracts meaningful latent representations of both users and articles. They used two real world social media datasets that have been used in previous works, Twitter and Weibo. Their work demonstrated the value in modeling the three intuitive and powerful characteristics of fake news.

(Ma, et al., 2015) proposed a novel approach to capture the temporal characteristics of these features based on the time series of rumor’s lifecycle, for which time series modeling technique is applied to incorporate various social context information. Their experiments using the events in two microblog datasets confirmed that the method outperforms state-of-the-art rumor detection approaches by large margins. Moreover, their model demonstrated strong performance on detecting rumors at early stage after their initial broadcast. They used linear SVM classifier for their model. They made comparison between DSTS-based SVM model and several strong baselines: DT: The Twitter credibility model using Decision Tree classifier with the social context features without considering time series; RF: The Random Forest classifier; RF-ext: Extension of the RF model; SVM-RBF: The SVM-based model with RBF kernel without considering time series; SV MDSTS c , SV MDSTS u , SV MDSTS d : DSTS model using content-based, user-based and diffusion-based features, respectively; SV MDSTS all : fully configured DSTS model. Overall, their system SV MDSTS, although a linear model, clearly outperformsed the baselines that were all based on non-linear models.

(Shu, et al., 2017) explored the correlations of publisher bias, news stance, and relevant user engagements simultaneously, and proposed a Tri-Relationship Fake News detection framework (TriFN). The Tri-relation stated by them is publisher, news and social engagements. They also provided two comprehensive real-world fake news datasets to facilitate fake news research. Datasets used include news from: BuzzFeedNews , LIAR, BS Detector and CREDBANK. The related social media posts were collected from Twitter using API by searching the headlines of news. TriFN can extract effective features from news publisher and user engagements separately, as well as capture the interrelationship simultaneously.

(Karimi, et al., 2018) introduced approaches to combine information from multiple sources and to discriminate between different degrees of fakeness, and propose a Multi-source Multi-class Fake news Detection framework MMFD, which combined automated feature extraction, multi-source fusion and automated degrees of fakeness detection into a coherent and interpretable model. They proposed a deep model to extract features from textual sources based on the CNN (convolutional Neural Network) and the LSTM (Long Short-Term Memory) network. They demonstrated that their model can effectively distinguish different degrees of the fakeness of news. In fakeness discrimination, they treated all classes the same.

(Shu, et al., 2018) constructed real-world datasets measuring users trust level on fake news and selected representative groups of both “experienced” users who were able to recognize fake news items as false and naive users who were more likely to believe fake news. They performed a comparative analysis over explicit and implicit profile features between those user groups, which revealed their potential to differentiate fake news. The findings of this paper laid the foundation for future automatic fake news detection research. Datasets were collected from PolitiFact and BuzzFeed. They identified the user groups by analyzing the user-news interactions, and computed the number of fake (real) news that user has shared. Then they proposed to use another metric, named Fake News Ratio (FNR). They then characterized user profiles by collecting and analyzing user profile features from different aspects, i.e., explicit and implicit. Experimental results on real-world datasets demonstrated that there are specific users who are more likely to trust fake news than real news; and these users reveal different features from those who are more likely to trust real news. These observations ease the feature construction of profiles features for fake news detection.

(Long, et al., 2017) proposed a novel method to incorporate speaker profiles into an attention based LSTM model for fake news detection. Speaker profiles contributed to the model in two ways. One was to include them in the attention model. The other included them as additional input data. By adding speaker profiles such as party affiliation, speaker title, location and credit history, their model outperformed the state-of-the-art method by 14.5% in accuracy using a benchmark fake news detection dataset. This proved that speaker profiles provide valuable information to validate the credibility of news articles. Experimental results showed that both methods of using speaker profiles can contribute to the improvement of fake news detection. This can be interpreted as speaker’s intention to speak the truth or fake it largely depends on his/her, profiles, especially his/hers credit history.

(Yang, et al., 2018) proposed a model named as TI-CNN (Text and Image information based Convolutional Neural Network). By projecting the explicit and latent features into a unified feature space, TI-CNN is trained with both the text and image information simultaneously. Extensive experiments carried on the real-world fake news datasets have demonstrated the effectiveness of TI-CNN in solving the fake new detection problem. The dataset in this paper contained 20,015 news, i.e., 11,941 fake news and 8,074 real news. It contained text and metadata scraped from more than 240 websites. Text Analysis included Computational Linguistic. Cognitive perspective, Psychology perspective, Lexical Diversity, Image Analysis and Sentiment Analysis were performed. Results showed that besides, the convolutional neural network makes the model to see the entire input at once, and it can be trained much faster than LSTM and many other RNN models.

(DiFranzo & Gloria-Garcia, 2017) stated that social media constructs “filter bubbles,” digital echo chambers where users see content and posts that agree only with their preexisting beliefs. They highlighted a work that explored whether it contributed to the 2016 US Presidential election results. According to them Personal recommendation systems, or systems that learn and react to individual users, have been claimed to be one cause of filter bubbles.

(Che, et al., 2018) examined the 10 popular partisan media sites discuss “fake news”. They used linguistic analysis techniques including Linguistic Inquiry and Word Count (LIWC), word embedding models, and supervised learning classifiers to analyze news stories containing the phrase “fake news” from left- and right-leaning news sites. They trained a logistic regression classifier to predict the political slant (left- or right-leaning) of news articles with the phrase “fake news”. Their results yielded several insights, including that article text can be used to classify political affiliation with high accuracy, and that left-leaning sites focus on specific fake news stories and individuals involved, while right-leaning sites shift the focus to a narrative of mainstream media dishonesty more broadly. The results underscored the politically polarized nature of partisan news sources, as well as the two parallel conversations—one about conspiracy theory-like examples and the other about media trustworthiness—which surround the topic of fake news.

(Xu, et al., 2018) characterized hundreds of popular fake and real news measured by shares, reactions, and comments on Facebook from two perspectives: Web sites and content. Their site analysis reveals that the Web sites of the fake and real news publishers exhibit diverse registration behaviors and registration timing. In addition, fake news tends to disappear from the Web after a certain amount of time. The content characterizations on the fake and real news corpus suggested that simply applying term frequency - inverse document frequency (tf-idf) and Latent Dirichlet allocation (LDA) topic modeling is inefficient in detecting fake news, while exploring document similarity with the term and word vectors is a very promising direction for predicting fake and real news.

(Parikh & Atrey, 2018) presented an insight on characterization of news story in the modern diaspora combined with the differential content types of news story and its impact on readers. They dived into existing fake news detection approaches that are heavily based on text-based analysis, and also described popular fake news data-sets. They used various Linguistic Features based Methods such as Ngrams, Punctuation, Psycho-linguistic features, Readability and Syntax.

They also used various Deception Modeling based Methods such as RST and VSM. RST-VSM methodology gave them an edge of curating data much better than similarity cluster analysis, which is simply based on distances between samples in the original vector space. They then used Clustering based Methods, Predictive Modeling based Methods, Content Cues based Methods and Non-Text Cues based Methods. They concluded their research by identifying 4 key open research challenges that can guide future research which were Multi-modal Data-set, Multi-modal Verification Method, Source Verification and Author Credibility Check.

(Buntain & Golbeck, 2017) developed a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowdsourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. They applied this method to Twitter content sourced from BuzzFeed’s fake news dataset. They used several features to predict fake news such as Structural Features, User Features, Content Features and Temporal Features. Results showed models trained against crowdsourced workers outperform models based on journalists’ assessment and models trained on a pooled dataset of both crowdsourced workers and journalists.

(Sirajudeen, et al., 2017) proposed approach which is a multi-layered evaluations technique to be built as an app, where all information read online is associated with a tag, given a description of the facts about the contain. They used three difference phases and dwells on a proof of concept for detecting an online fake news. The first phase is proposing and designing of a conceptual framework based on source and content of information that makes of a fake news. This framework has been designed to show how to combat the spread of fake news. The next step is a formulation of an algorithms necessary to implement the propose framework. The key algorithms involve; detection and filtering of online fake news. Hence, a step-by-step phase of detection and filtering of online fake news given. Finally, the last step is the development of a proof of the concept using Java programming language. Multi-layered evaluations contain : Verifying IP Address, Detecting the Source of Fake News, Determining the Status of News and Filtering of News. They provided steps needed to be taken by some Microblogging sites to prevent the dissemination of online fake news.

(Shi & Weninger, 2016) stated that Traditional fact checking by experts and analysts cannot keep pace with the volume of newly created information. It is important and necessary, therefore, to enhance the ability to computationally determine whether some statement of fact is true or false. They viewed this problem as a link-prediction task in a knowledge graph, and showed that a new model of the top discriminative meta paths is able to understand the meaning of some statement and accurately determine its veracity. They evaluate their approach by examining thousands of claims related to history, geography, biology, and politics using public, million node knowledge graphs extracted from Wikipedia and SemMedDB. Their approach significantly outperformed related models, they also found out that the discriminative path model is easily interpretable and provides sensible reasons for the final determination.

(Sherif, et al., 2017) introduced a new n-gram model to detect automatically fake contents with a particular focus on fake reviews and fake news. They studied and compared 2 different features extraction techniques and 6 machine learning classification techniques namely, stochastic gradient descent (SGD), SVM, linear support vector machines (LSVM), K-nearest neighbor (KNN), LR, and decision trees (DT). Experimental evaluation using existing public datasets and a newly introduced fake news dataset indicated very encouraging and improved performances compared to the state-of-the-art methods. They obtained 87%accuracy using n-gram features and the LSVM algorithm when classifying fake news against real news.

# Research Gap

Related research on detection of fake news suggested that the researchers were focusing on a single stream of data such as social media or newspapers/blogs or dating websites in terms of source. Secondly, the amount of data that was considered for research, analysis and predicting purposes was not sufficient. Thirdly, not every scenario can be analyzed using a similar analytical approach. There are limitations to the algorithms with respect to data content, size of data and application. Based on the 3 identified gaps, our research is grounded on detecting falsification in social media, news and blogs as a whole and adopting proper analytical methodology suitable for each of the scenario and to gather more data in order to identify provable patterns of linguistic leakages and thereby classifying the content as valid/reliable and fake.

# Objectives

* Collecting news articles from 3 different sources – News websites, social networks and blogs.
* Data preprocessing in order to make the data fit to apply machine learning algorithms.
* Try to identify word patterns in real vs fake news.
* Train multiple models which will be able to predict fake news from 3 different sources together.

# Research Design

The main area of focus for this paper is to perform a qualitative research to detect the authenticity of the news.

## **Data Collection and Scaling:**

The data used for this paper will be drawn from two different sources, both in public domain.

1. Collecting Legitimate News: We started with collection of a dataset of legitimate news belonging to business, entertainment and politics domains. The news obtained from a variety of mainstream news websites such as BBC, NDTV, and The Times of India etc. To ensure the reliability of the news, we conducted a manual fact-checking (Shi & Weninger, 2016) on the news content, like verifying the news source and cross-referencing information among several sources. Using this approach, we collected 50 news in each of the three domains, for a total of 150 authentic news.
2. Collecting Fake News: Similarly we collected the fake news from web sources following similar guidelines as in the previous dataset. However, this time, we aimed to identify fake content that naturally occurs on the web along with the Kaggle data sets with 50 each in business, entertainment and politics domains.

Using this approach, we collected 150 fake news articles and 150 legitimate news articles in the business, entertainment and politics domains. As our data size was less, we also collected dataset from Kaggle which had 6035 articles.

## 

## **Research Methodology**

This segment delivers a brief impression of the methodology to analyze the collected data. This section describe the several different models which can be implemented:

1. Pre-processing the Text: We used the packages in Python to tokenize our text and performed the steps of removing stop-words (Sherif, et al., 2017) like “some”, “there” etc., followed by removing whitespace & Punctuation i.e. “\n”, “@” etc., then lemmatization of words which converted each word to its root form, turning different words into a single representation and last performed N-grams to combine nearby words into single features (Ahmed, et al., 2017), which helped to give context to words that may have little meaning on their own.
2. Converting Text to Features: To analyze and model text after it has been preprocessed, it must first be converted into features. Techniques may include TF-IDF or Word2Vec (Ahmed, et al., 2017).
3. Models: We utilized Naïve Bayes (Aphiwongsophon & Chongstitvatana, 2018) , Logistic Regression (Ahmed, et al., 2017) and Random Forest (Seale, 2018) for the implementation part. We performed a cross-validation (Ahmed, et al., 2017) and a grid search (Jamrul, 2017) to find the best parameters for the TF-IDF algorithm and each individual model.

# Fake news

Types of fake news

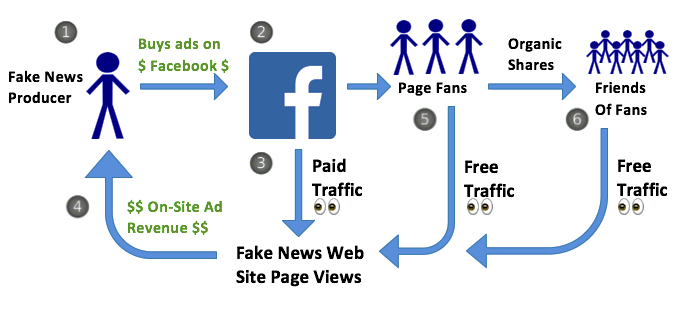
There are 5 different types of fake news as classified below based on its nature.



## 

## Spread of fake news

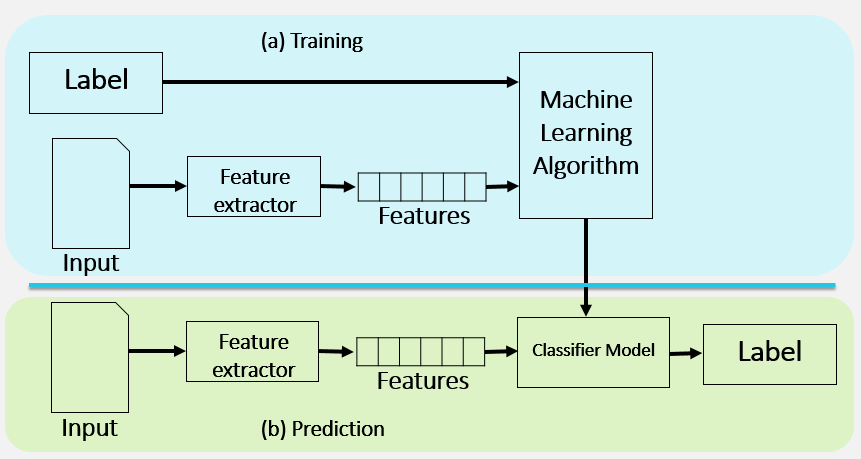
Below picture depicts the manner in which the fake news is spread from the producer to the masses.



www.freedom-to-tinker.com

# Modelling approach

The below schematic diagram explains the method in which we have trained our algorithm by feeding the processed news articles (text) as input. It also explains how we used the model to predict the newer articles.



Steps taken to apply the model:

## Pre-Processing the Text

* Combined data of ‘Title’ and ‘Text’ (Vedova, et al., 2018)
* Imported Stop words from Natural Language Toolkit Library of Python
* Stop words include”I me when where both few should” etc.
* Removed all words having length less than 3 and greater than 15
* Converted all the words to lower case
* Removed all the characters except (a-z)
* Split data into train and test set with 4751 articles as training set and 1584 articles as test set.

## Converting Text to Features

* Used CountVectorizer from SKLEARN Library (Sherif, et al., 2017)
* Created vocabulary from training data with the use of fit() function.
* Created Encoded Vector with the help of transform() function on training data.

Encoded vector returned length of entire vocabulary and an integer count for the number of times each word appeared in the document.

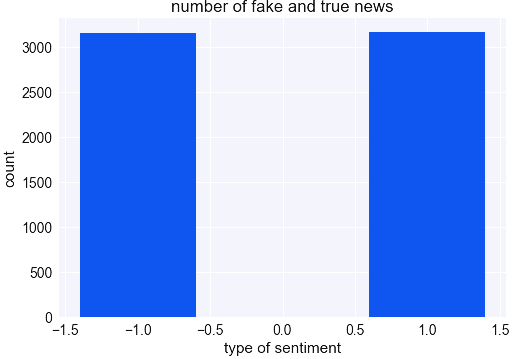
## 

## Models Applied

* Naïve Bayes
* Logistic Regression
* Random Forest

**Exploratory data analysis**

The below diagram shows the number of fake and true news in the crowdsourced corpus of news articles. Number of fake news articles are 3164 and real news articles are 3171.



Highest number of real and fake tokens found in articles:

|  |  |
| --- | --- |
| Real Tokens | Count |
| Party | 2756 |
| House | 2736 |
| Sanders | 2648 |
| Percent | 2611 |

|  |  |
| --- | --- |
| Fake Tokens | Count |
| World | 2234 |
| Government | 1984 |
| Obama | 1526 |
| Russia | 1432 |

# Model performance

A total of 3 different machine learning algorithms were used. Below is the summary of performance of the models on test data.

## Naïve Bayes

Accuracy on training data = 0.9444327509997895

Accuracy on test data = 0.8869949494949495

Confusion Matrix = [[681 136]

[ 43 724]]

In confusion matrix, value 681 shows True Negatives and value 724 shows True Positives. So, most of the news articles are correctly classified and so we got such a high accuracy.

## Logistic Regression

Accuracy on training data = 0.959834567374834

Accuracy on test data = 0.9065656565656566

Confusion Matrix = [[743 74]

[ 74 693]]

In confusion matrix, value 743 shows True Negative value and value 693 shows True Positive

## Random Forest

Accuracy on training data = 0.9997895179962113

Accuracy on test data = 0.8693181818181818

Accuracy on training data is nearly 1 which shows overfitting of model. (Vegt, 2018)

# Conclusion

Logistic Regression gives the accuracy of 90% on test data set which is best among all 3 models. Hence, Logistic Regression is found to be the most appropriate classification model to detect fake news on social media platforms.

# Limitations

Below are the observed limitations of the project:

* Number of observations are less and corpus of words/vocabulary is created only on the basis of these articles.
* Model might not perform properly in case of news articles which doesn’t contain the words on which it is trained.
* Only 180 common ‘stop words’ were removed, there can be many other words which don’t contribute much to classification of article as fake or real.
* Model works only on ‘proper’ English language articles.

# Future scope

* Identifying ‘Linguistic cues to deception’
* Implement Natural Language Processing in combination with Neural Networks
* Extracting punctuation patterns and ‘N-grams’ in text
* Use Rhetorical Structure Theory to identify rhetoric relations between linguistic cues