2018 IEEE International Conference on Information Reuse and Integration for Data Science

Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers

Opeyemi Mulikat Aborisade
Department of Applied Mathematics
North Carolina A& T State University
Greensboro, USA
Email: omaborisade@aggies.ncat.edu

Mohd Anwar
Department of Computer Science
North Carolina A& T State University
Greensboro, USA
Email: manwar@ncat.edu

***Abstract*—At a time when all it takes to open a Twitter account is a mobile phone, the act of authenticating information encountered on social media becomes very complex, especially when we lack measures to verify digital identities in the first place. Because the platform supports anonymity, fake news generated by dubious sources have been observed to travel much faster and farther than real news. Hence, we need valid measures to identify authors of misinformation to avert these consequences. Researchers propose different authorship attribution techniques to approach this kind of problem. However, because tweets are made up of only 280 characters, finding a suitable authorship attribution technique is a challenge.**

This research aims to classify authors of tweets by comparing machine learning methods like logistic regression and naive Bayes. The processes of this application are fetching of tweets, pre-processing, feature extraction, and developing a machine learning model for classification. This paper illustrates the text classification for authorship process using machine learning techniques. In total, there were 46,895 tweets used as both training and testing data, and unique features specific to Twitter were extracted. Several steps were done in the pre-processing phase, including removal of short texts, removal of stop-words and punctuations, tokenizing and stemming of texts as well. This approach transforms the pre-processed data into a set of feature vector in Python. Logistic regression and naive Bayes algorithms were applied to the set of feature vectors for the training and testing of the classifier. The logistic regression based classifier gave the highest accuracy of 91.1% compared to the naive Bayes classifier with 89.8%.

***Keywords*-authorship attribution; social computing; privacy; security; machine learning; logistic regression; naive Bayes; classification**

I. Introduction

With the growth and popularity of online social media, the participation in online social platform has become essential for individuals and organizations to remain competitive. Governments and powerful individuals have used information as a weapon, to boost their support and quash dissidence [12]. In a recent fake news inquiry in UK, Twitter and Facebook were included in the names of companies that were threatened with sanctions, which resulted in people questioning how Twitter can help identify known sources of misinformation [1].

An online social network (OSN) is an online platform of websites and applications dedicated to facilitate social interactions and personal relationships. The users of OSN like Facebook, Twitter, Instagram, Snapchat, etc. can also involve in malicious activities like the spreading of fake news with no reliable sources. Anonymous Social Networks (ASN) like Whisper, Secret, Wut, and Yik Yak, provide the same service as the OSN to the users but encourage the users to stay anonymous or adopt pseudonyms [23]. The use of social media has become a part of people's daily activities, but it can cause significant harm by propagating misinformation. Therefore, determining authorship or classifying authors to source of information is a prevalent issue on social networks.

Twitter, a social network application, is ranked fourth most popular OSN in the United States of America. Twitter has about 330 million monthly active users with about 500 millions of tweets per day. The author of Politics and the Twitter Revolution, John Parmelee said Twitter can set the agenda for what journalists are covering [23]. Twitter provides users the freedom to communicate, collaborate, network and share various information regardless of who you are. A tweet is a 280-character message that can include opinion or information about recent happenings including users reactions. A tweet consists of pictures, videos, hash-tags, hyper-links, emojis, locations, gifs, and polls. Classifying tweets to their authors is a crucial need. Researchers suggest the use of authorship attribution.

Authorship attribution is a process of identifying author of a given corpora, given the collection of corpora whose authorship is known. It is the way of deducing who the author of a given text is, when it is unclear who wrote the text. As anonymous information increases with rapid increase of internet usage around the world, authorship attribution has become a valuable tool. Some of the common applications of authorship attribution include plagiarism detection, deducing the writer of threatening emails, identifying authors of fake news. It is useful in settling dispute between two people who had claim to be

978-1-5386-2659-7/18/$31.00 ©2018 IEEE
DOI 10.1109/IRI.2018.00049

the author of a given corpora [18]. Authorship attribution is useful for text classification with emphasis on the text content. Authorship attribution can be used in a broad range of applications in diverse areas, including intelligence, criminal and civil law, computer forensics, and cyber-crime investigation as well as in the traditional application to literary research.

It is important to be able to identify the author of tweets since this will raise consciousness and decrease any possible propagation of misinformation. Some people can imitate someone else's style of writing as well. The challenge in authorship attribution is when you find a person who publicly posts in multiple formats, but their voice radically changes depending on their audience [5]. This is what makes authorship attribution a prevalent issue and due to the limit of characters in a tweet, using tweets for authorship attribution adds a layer of complexity.

This research, therefore, aims to build a model using supervised learning technique using logistic regression and naive Bayes methods to classify authors of a given corpora within the social network, Twitter. we explore classification of authors to it's tweets with focus on high quality training texts. Machine learning can be used to provide computational intelligence to the technology, so it can learn and adapt from the given data independently [25]. We used the combination of natural language processing toolkit and Scikit-learn package in Python. Since the dataset is made up of two target classes (known authors and unknown authors), we will focus on binary classification.

This research is divided into sections, with the just concluded section as the section one. Section two presents related works on authorship attribution. Section three is the framework/approach of the research. It has four main phases which are data collection, data pre-processing, feature extraction from tweets and machine learning application. Section four is the results, and conclusion section is the fifth section.

II. RELATED WORK

Many new methods have been developed in recent years to address the issue of authorship attribution. There has been a great amount of work done on authorship attribution of unstructured or structured text. When anonymity is being discussed, there are two center of attraction; either to reveal authorship or conceal it. Rather than concealing authorship, Qian et al. (2016) designed a three-way classification system to reveal authorship of a corpora. They developed a Tri-Training Algorithm as well as other strategies such as Inter and Self Adding to make authorship known. This approach improve upon the CNG+SVM method that used just two classifiers in determining authorship attribution [20].

On the other hand, McDonald et al. (2012) considered the opinion that anonymity of authorship within a corpus should be concealed for privacy and security sake. They suggested that adversaries of those that produce certain importance to internet-based literature may attempt to use known authorship attribution and stylometry techniques to unveil the authorship of said literature. They introduced two open source tools called JStylo and Anonymouth. The tools create anonymity for documents that are pre-existing to preserve author's anonymity [16].

Progressively, researchers are finding their way to resolve authorship attribution without the use of known stylometric technique. Koppel et al. (2006) worked on authorship attribution with thousands of blog authors. They used meta learning which involves using three text categorization method to run their experiment as well as similarity scores to complete authorship attribution. The focus of the article was to show that authorship attribution can be solved to a reasonable extent even when there are thousands of candidate authors [15]. Schimid et al. (2015) also worked on email authorship using customized associative classification. They opted to use data mining techniques instead of the normal stylometric or attribution techniques. Their findings show that the more features you have, the more promising your results will be [21].

Rocha et al. (2017) discussed authorship attribution techniques that can be used to detect authorship for the purpose of social media forensics. The utmost challenge faced with social media texts is the length of its text characters like we have in Twitter. They examine supervised learning based methods that are effective for small sample sizes. It was also argued in this paper that there is a need for new authorship attribution applications that can use context and process data obtained from multiple modalities. The benefits of developing automated applications for authorship attribution will aid researchers and investigators to easily determine and eliminate suspects for cybercrime [2].

Research in the past had focused on revealing/concealing authorship and proposing new classification method, which were evaluated over some commonly used text corpuses, and its performance was compared with that of the existing methods. There is a need for a new application for authorship with focus on how the training corpus helps to impact classification performance. To solve the aforesaid challenges, we explore classification of authors to their tweets with focus on high quality training texts. There are implications to this research which will be discussed later in the conclusion and recommendation section.

III. FRAMEWORK

In this section, we discussed the text classification process, which includes data collection, data pre-processing, feature extraction and machine learning application. This steps can be illustrated as seen in *fig 1*:

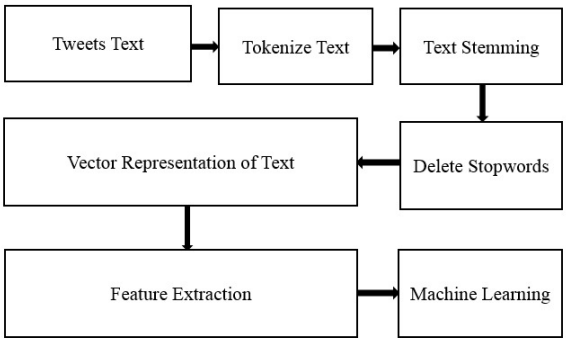

Figure 1. Text Classification Process [27]

A. Dataset

The dataset used for this research was streamed from Twitter, with maximum of three thousand tweets per author. The dataset consist of two features, the author (user name) and the tweets itself. The author is the response variable, which is made up of two classes, known and unknown authors. The known authors consist of twelve celebrities and prominent users on Twitter. While the unknown authors also consist of twelve regular Twitter users, but not as popular as the known authors. The tweets were labeled known and unknown based on the author's class as a celebrity or not. We defined celebrities on Twitter as prominent people in the field of politics and media journalism.

B. Data Collection

We collected our data directly from Twitter server using the Twitter API (Application Programming Interface) based on user ID. The Twitter API as over the time been split into multiple APIs such as Twitter Ads API, Twitter Search Tweets API, and Twitter Direct Message API. The Twitter micro-blogging service includes two RESTful APIs. The Twitter RESTful API methods allow users to access important Twitter data which includes status data, timeline and user information. The Search API methods give users methods to interact with Twitter Search and trends data [28].

For this research, we used the RESTful API for fetching of tweets since the focus is on the user's tweets. There are few steps in the data collection phase, including installation of tweepy (a Python Library to easily access the Twitter API), creating a developer's account to have access to tokens, creating an instance of a tweepy to handle the incoming data, collect tweets on user ID, transform the collected tweets into an array that will populate the csv and write the csv document.

The dataset consists of the author and the tweets only, all re-tweets which are messages retransmitted from other users, were removed since the goal of the research is on authorship attribution. The Twitter dataset used for this research is made up of 46,895 tweets with 22,333 tweets in the known class and 24,566 in the unknown class of authors. The known authors consist of politicians, celebrities and prominent people on Twitter while the unknown authors are regular Twitter users.

C. Data Pre-processing

Data Pre-processing is an approach of removing unwanted terms like hyper-links, stop words, and outliers that do not add value to the classification of texts [25]. When we want to build Machine Learning model based on tweets, a pre-processing of the tweet is required. We build methods in Python language, which support cleaning, normalizing, tokenizing and stemming of texts. The stages involved were:

1) Removal of short tweets that does not give useful information about the author.
2) Normalizing of hyper links, re-tweeted tweets and hashtags with a new text URL, RT, and TAG respectively.
3) Removal of stop words like *a, is, and, most,* and so on in the text because those words are commonly used and are not unique to each author.
4) Tokenizing the texts to break text into words.

D. Feature Extraction

Feature extraction is a process of converting texts to vector form that will be used as the input. In other words, in feature extraction we begin by creating a map from words to a vector space. There are a lot of features for authorship attribution that have been proposed in the literature over time [4]. For this research, we considered the Bag of Word model and Vector Space Model - Term Frequency-Inverse Document Frequency for feature extraction.

Bag of Words Model: Bag of Word (BoW) model involves transforming all the texts into a dictionary that consist of all words that appear in all texts [2]. BoW model is frequently used in text classification. BoW involves classifying and analysing different corpus by taking in a set of documents as an input and gives a table with frequency counts of each words in each documents. Let's take for example the following texts [2]:

Text 1: "I love my Boy"
Text 2: "My Boy is awesome"

From this texts, we have the dictionary that maps each feature to unique index, which is most times referred to as document frequency (df) as illustrated in *Table I.*

Word	i	love	my	boy	is	awesome
df	1	1	2	2	1	1

Table I: DOCUMENT FREQUENCY DICTIONARY FOR A BAG OF WORD MODEL DERIVED FROM TWO LINES OF TEXT INSPIRED BY [16]

The word for each line of texts are used as the features. From the dictionary, we can generate separate feature vectors by counting the words in each documents.

Term Frequency-Inverse Document Frequency: Other statistics such as the Term Frequency-Inverse Document Frequency (TF-IDF) scores can be considered in text classification. The TF-IDF document frequency compute frequency that a word appears in a document compared to its frequency across all documents [10]. The term frequency (tf) of the two lines of text example given above is shown in *Table II* and the associated TF-IDF, which is the ratio of tf and df is given below in *Table III* respectively. We have used the TF-IDF library in Python for this research.

Word	i	love	my	boy	is	awesome
Text 1	1	1	1	1	0	1
Text 2	0	0	1	1	1	1

Table II: TERM FREQUENCY DICTIONARY FOR A BAG OF WORD MODEL DERIVED FROM TWO LINES OF TEXT INSPIRED BY [16]

Word	i	love	my	boy	is	awesome
Text 1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	0	1
Text 2	0	0	$\frac{1}{2}$	$\frac{1}{2}$	1	1

Table III: TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY DICTIONARY FOR A BAG OF WORD MODEL DERIVED FROM TWO LINES OF TEXT INSPIRED BY [16]

E. Machine Learning

After feature extraction, the documents can be easily represented in a form that can be used by a machine learning algorithm. A great number of text classifiers have been proposed in the literature using machine learning techniques. Depending on the approach adopted, the machine learning techniques differ [17]. Although many approaches have been proposed, text classification is still a major area of research that needs improvement.
In text classification application, we use naive Bayes often because of its clarity and effectiveness [13]. However, its performance is often regarded with contempt because it does not model text perfectly. This problem was addressed by Schneider and it being shown that it can be solved by some simple corrections [22]. Support vector machines (SVM), when applied to text classification provide excellent precision, but poor recall. One way to customize SVM's to improve recall, is to adjust the threshold associated with an SVM as described by Shanahan and Roma [24].
The previous research like real time tweets detection for small scale incidents [3] and tweets classification for alcohol use [30] revealed that logistic regression can generate better results in classifying text than any other methods like naive Bayes, decision tree, and other machine learning techniques.

1) Logistic Regression: Logistic regression is an advanced Linear regression technique used for classifying both linear and non-linear data. It is commonly used to model data with binary responses. When the response is binary, it takes the form 0/1, with 1 generally indicating a success and 0 a failure. But, the actual values that 1 and 0 can take vary widely, depending on the goal of the study. For this research, authors are classified to be either known or unknown, where 1 represent known, and 0 for unknown authors.
Logistic regression is a machine learning technique that is implemented by taking the given input value and multiply the input with weight value [25]. Consider a document (X) and class (C), logistic regression directly estimates the parameters of $P(C|X)$. For the tweets dataset considered in this research, logistic regression models the probability of authorship. Mathematically, we have

$$Pr(class \quad of \quad authors|Tweets) = Pr(Tweets)$$

where the author response falls into one of two class categories, known or unknown. That is, r which is the possible values of class C is given as

$$r = \begin{cases} 0 & known \\ 1 & unknown \end{cases}$$

a) *Computation of the Model*
Logistic regression is often referred to as a discriminative classifier. That is, its an approach to learning functions of the form $f : X \rightarrow C$, or $P(C|X)$ in the case where C is discrete-value, and $X = x_1 \cdot \vec{\cdot} x_N$ is a vector containing discrete or continuous variables [26]. The equation for this is as shown below:

$$P(C|X) = \sum_{i=1}^{N} w_i^{\top} x_i \tag{1}$$

The values of $P(C|X)$ cannot be calculated directly using the previous equation because it gives values that range from $-\infty$ to ∞. But we want to be able to generate a value of an output that ranges between 0 and 1, hence we use the following exponential function [25]:

$$P(C|X) = \frac{1}{z} e^{w^{\top} x} \tag{2}$$

Changing the normalization factor z, and specifying the number of features (N), we have as follows:

$$P(C|X) = \frac{\exp(\sum_{i=1}^{N} w_i^\top x_i)}{\sum_r \exp(\sum_{i=1}^{N} w_i^\top x_i)} \tag{3}$$

b) *Model Fitting*

To fit our model, we use the method called Conditional Maximum Likelihood (CML). CML is used by logistic regression as the estimator of the weight (w) value. For this method to work, we choose a value of w that maximizes the probable value of class C given the input X. The CML equation is given below as [25]:

$$\hat{w} = argmax_w \sum_j \log P(r^j|x^j) \tag{4}$$

$$L(w^\top) = \sum_j \log P(r^j|x^j) \tag{5}$$

$$L(w^\top) = \log \sum_j \exp(\sum_{i=1}^{N} w_i^\top x_i(r^{(j)}, x^{(i=j)}) \tag{6}$$

$$- \log \sum_j \sum_{r \in C} \exp(\sum_{i=1}^{N} w_i^\top x_i(r^{(j)}, x^{(i=j)}))$$

2) Naive Bayes: Naive Bayes is a supervised machine learning classifier that is based of Bayes rule. Naive Bayes classifier is one of the most useful supervised learning algorithm for text classification [6]. Consider a document (X) and class (C), naive Bayes classifier directly estimates the parameters of $P(C)$ and $P(X|C)$. The Bayes rule is of the form:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{7}$$

The task is to train the model to predict the correct class for a new instance. Following the Bayes rule, we can compute the *maximum a posteriori* class (the most likely class) for the given data, which is given below: [9]

$$C_{map} = arg \max_{r \in C} P(X|r)P(r) \tag{8}$$

where C is the set of all class target. When features are extracted from the document, the document can be represented as set of features $(x_1 \cdots x_n)$. The maximum a posteriori class can be written as [6]:

$$C_{map} = arg \max_{r \in C} P(x_1 \cdots x_n|r)P(r) \tag{9}$$

However, to estimate $P(C)$, we then compute the relative frequency of each target class in the training data. Estimating $P(x_1, x_2 \cdots x_n|C)P(C)$ is tedious because there are not enough instances for each attribute combination in the training set, this lead to sparse data problem. Using the independence assumption that attributes are conditionally independent given the target class value. This can be expressed mathematically in the form [9]:

$$P(x_1, x_2 \cdots x_n|C)P(C) = \prod_i P(x_i|C) \tag{10}$$

Hence we get the following classifier

$$C_{NB} = arg \max_{r \in C} P(C) \prod_i P(x_i|C) \tag{11}$$

3) Confusion Matrix: Confusion Matrix is a table matrix that is often used for performance evaluation in classification. Performance in classification is mostly expressed in terms of accuracy. An example of confusion matrix for a binary classification task is given in *table IV*. From the confusion matrix, we can calculate the accuracy rate of the classifier.

	Predicted Known	Predicted Unknown
Actual known	TN	FP
Actual Unknown	FN	TP

Table IV: CONFUSION MATRIX EXAMPLE

IV. RESULTS

The Twitter dataset used for this research is made up of 46,895 tweets with 22,333 tweets in the known class and 24,566 in the unknown class of authors. The known authors consists of politicians, celebrities and prominent people on Twitter while the unknown authors are regular Twitter users. The dataset was uploaded to a SQLite database. This research used the train-test-split library in scikit-learn for splitting the dataset into training and testing set. The training set consist of 35312 tweets while the testing set has 8828 tweets.

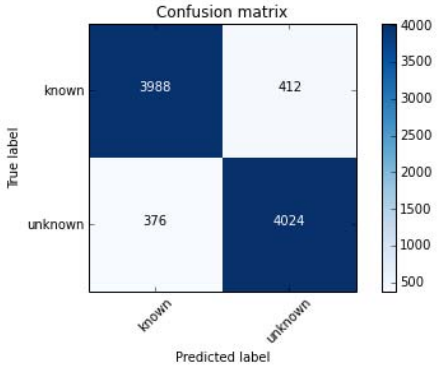

Figure 2. LOGISTIC REGRESSION MODEL CONFUSION MATRIX

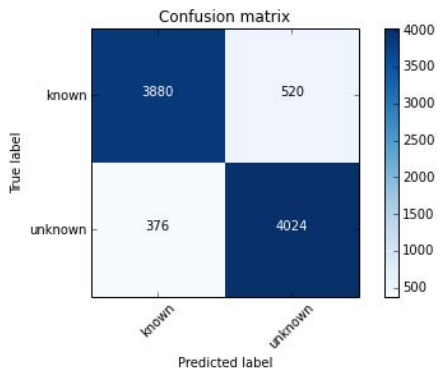

Figure 3. NAIVE BAYES MODEL CONFUSION MATRIX

For the logistic regression model, we have the accuracy computed from the confusion matrix to be;

$$Accuracy_{LR} = \frac{3988 + 4024}{8800} = 0.9105 \tag{12}$$

And for the naive Bayes model we have,

$$Accuracy_{NB} = \frac{3880 + 4024}{8800} = 0.8982 \tag{13}$$

Furthermore, we computed the accuracy in Python to verify the already computed accuracy. The best parameter resulted in a 91.1% accuracy for the logistic regression classifier, which is on the high side compare to the naive Bayes classifier with 89.8%. We noticed that the naive Bayes classifier seem faster than the logistic regression classifier with a computational time differences of 2.3mins.
From the result, we were able to classify tweets to its attributed authors with a higher accuracy. This approach is easy and flexible, and it was fully built in python.

Method	Best Score
Logistic Regression	0.9105
Naive Bayes	0.8982

Table V: SCORE SUMMARY BY MODEL

V. CONCLUSION AND RECOMMENDATION

A. CONCLUSION

In this research, we were able to collect tweets from Twitter using the RESTful Twitter API and pre-process the tweets for authorship attribution. We extracted features from the tweets using bag of words and vector space model which demonstrated good performance on tweets classification. After feature extraction, we used logistic regression and naive Bayes machine learning techniques for classification. The logistic regression classifier resulted in an accuracy greater than 91% accuracy and the Naive Bayes classifier gave a 90% accuracy as shown in *table v*.
Our contribution to text classification shows that classifiers performance is relevant to its training corpus in some extent, and having a well pre-processed training corpora may generate classifiers of good performance. Unfortunately, little research work has been seen over time on how to exploit training corpus's to help improve classifier's performance.
In literature, it has been observed that for specified classification method, classification performances of the classifier are based on different training text corpora. This means that the greater the quality of training corpora, the better the classifiers performance. Growing volumes and variety of available data in Twitter can be further classified to its attributed authors through iterative learning from the dataset because it will require little human intervention due to its ability to learn. It is hoped that this approach can be adopted for authorship attribution for later use.

B. RECOMMENDATION

We can further classify tweets to its authors by studying the methodology of feature extraction under concepts, to observe if these will help in text classification. It is well known in the literature that vector space models suffer greatly. One of it's limitation is the assumption of independence between terms. The fact that occurrences of a term says nothing about another term occurring is viewed as a limitation and the implication of this limitation are still in debate. Some important conclusions have not been reached yet, since this paper is just a starting point of our research in this direction. On the other hand, we will be considering sequence models for term dependencies.
Another limitation is being able to distinguish between different meanings of a word which researchers referred to as word sense disambiguation. Consideration of synonyms and phrasal relationships during feature extraction can lead to dimension reduction of the features for unique words that will aid better classification of tweets for authorship attribution. All aforementioned limitations lead to open questions that need urgent attention.
Authorship attribution can also be extended to other social media like Instagram, Facebook, making it easy to pin point authors for detection of malicious activities.

VI. ACKNOWLEDGEMENT

This research is based upon work supported by the United States Government including the National Science Foundation. Caleb kindly reviewed and made many suggestions that improved this final version. Siobahn and May made a great contribution through their opinions to the data collection section.

REFERENCES

[1] **Alex Hern,** ***"Facebook and Twitter threatened with sanctions in UK 'fake news' inquiry,"*** The guardian Dec 2017. [online].

[2] **A, Rocha, Walter J. Scheirer, Christopher W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, Authorship Attribution for Social Media Forensics**, IEEE Transactions on Information Forensics and Security (Volume: 12, Issue: 1, Jan. 2017).

[3] **A. Schulz, P. Ristoski dan H. Paulheim,** ***I See a Car Crash: Real-Time Detection of Small scale Incidents in Microblogs,*** dalam 10th Extended Semantic Web Conference, Montpellier.

[4] **Authorship Attribution with Python,** ***"Authorship Attribution with Python,"*** AICBT Data driven research and development. [online]. Available: http://www.aicbt.com/authorship-attribution/

[5] **Bruce Schneier,** ***"Identifying People by their Writing Style,"*** 3 August 2011. [online]. Available https://www.schneier.com/blog/archives/2011/08/identifying_peo_2.html.

[6] **C. D. Manning, Prabhakar Raghavan, Hinrich Schtze** ***"Naive bayes Text Classification,"*** 2008. [online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

[7] **D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz,** ***A decision-tree-based symbolic rule induction system for text categorization,*** IBM Systems Journal, September 2002.

[8] **Dreamgrow,** ***Top 15 Most Popular Social Networking Sites and Apps***, January 2018. [online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/

[9] **Frank Keller,** ***"Naive Bayes Classifiers; Connectionist and Statistical Language Processing,"*** Pg 7-17. [online]. Available: http://www2.cs.uh.edu/~arjun/courses/nlp/naive_bayes_keller.pdf

[10] **Gerardnico,***"Text Mining - term frequency - inverse document frequency (tf-idf),"* last modified 2017/03/17. [online]. Available: https://gerardnico.com/wiki/natural_language/tf-idf

[11] **How to find sources on Twitter,** ***"How to find sources on Twitter", An exercise:***, October 2015. [online]. Available: http://training.npr.org/social-media/how-to-find-sources-on-twitter-an-exercise/

[12] **J. Titcomb, J. Carson,** ***"Fake News: What exactly is it - and how can you spot it?"***, January 2018. [online]. Available: http://www.telegraph.co.uk/technology/0/fake-news-exactly-has-really-had-influence/

[13] **Kim S. B., Rim H. C., Yook D. S. and Lim H. S., LNAI 2417,** ***Effective Methods for Improving Naive Bayes Text Classifiers,*** Pacific Rim International Conference on Artificial Intelligence, 2002, pp. 414-423 https://link.springer.com/chapter/10.1007/3-540-45683-X_45

[14] **K. Markham, Data School,** ***Simple Guide to Confusion Matrix Terminology,*** 26 March 2014. [online] Available: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[15] **Koppel, M., Schler, J., Argamon, S., & Messeri, E.,** ***"Authorship attribution with thousands of candidate authors,"*** In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 659-660, August 2006. [online]. Available: https://www.researchgate.net/profile/Moshe_Koppel/publication/221300609_Authorship_attribution_with_thousands_of_candidate_authors/links/0912f50c5cf83a5e0a000000.pdf

[16] **McDonald, A. W., Afroz, S., Caliskan, A., Stolerman, A., & Greenstadt, R.,***"Use fewer instances of the letter "i": Toward writing style anonymization,"* In International Symposium on Privacy Enhancing Technologies Symposium, pp.299-318, 2012,July. [online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-31680-7_16

[17] **M. IKONOMAKIS , S. KOTSIANTIS, V. TAMPAKAS,** ***"Text Classification Using Machine Learning Techniques,"*** pp. 966-974, August 2005. [online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9153&rep=rep1&type=pdf

[18] **M. Sudheep Elayidom, Chinchu Jose, Anitta Puthussery, Neenu K Sasi,** ***"Text Classification for Authorship Attribution Analysis"***, September 2013.

[19] **Preprocessor,** ***"Elegant and Easy Tweet Preprocessing in Python,"*** updated 7 December 2017. [online]. Available: https://github.com/s/preprocessor

[20] **Qian, T., Liu, B., Chen, L., Peng, Z., Zhong, M., He, G., ⋯ & Xu, G.** ***"Tri-Training for authorship attribution with limited training data: a comprehensive study. Neurocomputing,"*** pg 171, 798-806, 2016. [online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215010711

[21] **Schmid, M. R., Iqbal, F., & Fung, B. C. (2015).** ***"E-mail authorship attribution using customized associative classification: Digital Investigation,"*** pg 14, S116-S126. [online]. Available: http://www.sciencedirect.com/science/article/pii/S1742287615000572

[22] **Schneider, K.,** ***"Techniques for Improving the Performance of Naive Bayes for Text Classification,"*** LNCS, Vol. 3406, 2005, 682-693

[23] **Secret List,** ***"The Secret List of Anonymous Social Networks ,"*** updated Feburary 2017. [online]. Available: https://www.lifewire.com/list-of-anonymous-social-networks-2654864

[24] **Shanahan J. and Roma N.,** ***"Improving SVM Text Classification Performance through Threshold Adjustment,"*** European Conference on Machine Learning LNAI 2837, 2003, 361-372 https://link.springer.com/chapter/10.1007/978-3-540-39857-8_33

[25] **S. T. Indra, L. Wikarsa, R. Turang,** ***"Using logistic regression method to classify tweets into the selected topics,"*** Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference http://ieeexplore.ieee.org/document/7872727/

[26] **Tom M. Mitchell,** ***"GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION of Machine Learning textbook ,"*** 2015.

[27] **T. Sileo, *"Using Twitter REST API,"* v1.1 with Python**, updated July 2013. [online]. Available: https://thomassileo.name/blog/2013/01/25/using-twitter-rest-api-v1-dot-1-with-python/

[28] **Twitter API, *"Twitter API"* 2017**. [online]. Available: https://www.programmableweb.com/api/twitter

[29] **Twitter character, *"Twitter officially expands its character count to 280 starting today"***: November 2017. [online]. Available: https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/

[30] **Y. Aphinyanaphongs, B. Ray, A. Statnikov dan P. Krebs, *Text Classification for Automatic Detection of Alcohol Use-Related Tweets,*** dalam Information Reuse and Integretion (IRI), 2014, IEEE 15th International Conference, Redwood City, 2014.