A Social Media Text Analysis Technique Using ML Algorithm

Prof. Pallavi Bagde
Computer science engineering dept.
Sushila devi Bansal College,Indore
MP ,India
pallubagde94@gmail.com

***Abstract*— Data mining and ML approaches are used for analyzing the raw data. The analysis consequences has depends on the applications requirements. The use of ML has also been become popular in analyzing the social media data based on NLP for uncovering the sentiments on the social media text post. Social media has two sides brighten and dark. There are a number of creative ways to use the social media, but there are also some users that are accomplishing their toxic intensions using social media. The use of machine learning algorithms with the social media helps us to performing prediction, classification, categorization and finding relationship among social media data attributes. Therefore, in this paper we have exploring the difference among the social media data text mining technique based data analysis and also by using NLP based text classification. In this context a publically available social media dataset from Kaggle has been collected and a data model has been presented to classify social media text using support vector machine (SVM) and backpropagation neural network (BPN) classifiers. In order to extract the features we have used the TF-IDF and in second scenario we have used the Part of speech (POS) tagger. The obtained results demonstrate the performance of BPN based classifier is higher in both the scenarios of feature classification. Additionally in simple subject classification the TF-IDF based features providing more yield as compared to POS based features.**

Keywords—machine learning algorithm, supervised learning, social media data, NLP based features, classification performance.

I. INTRODUCTION

The social media has become a popular platform for expending the time among all the age group of users. The people are usage social media for chatting and communication with their friends and loved ones, sharing the knowledge and new thoughts, providing review or opinion and can do many more things on social media. In addition, the social media has also been used for providing information, promotions, advertisements and others. Therefore the social media platforms are beneficial for social as well as professional tasks. The social media thus a huge collection of data and also contains a rich volume of opportunities for gaining knowledge and utilizing them in various social welfare applications. In this presented work we are proposed to use and investigate the social media data and use of machine learning together for finding new opportunities in social media data.

In this paper thus we have include the investigation of recent techniques for social media data analysis. Thus we introduce a review in next section. In this review we have included the different areas of applications and techniques utilized for designing social media data based applications. Further, we proposed a social media data analysis technique using the ML algorithms. The aim of this data modeling is to provide the understanding about the steps involved in social media text analysis. Additionally provide a comparative performance study among two popular feature selection techniques and the ML based classification techniques for social media text post classification.

Therefore the experiment includes the explanation and implementation of Term Frequency- Inverted Document Frequency (TF-IDF) and part of Speech (POS) based feature selection techniques. Additionally to classify these features we have used the support vector machine (SVM) and backpropagation neural network (BPN). these developed algorithms have been used with the twitter social media dataset for classifying the negative, positive and neutral orientation of text. The experiments has been carried out, and based on the experimental observations we have provides the results analysis and their comparative performance. Finally based on the conducted efforts we have provided some lighting insights and the future research objectives to be accomplished.

II. RELATED WORK

In this section we provide the basic information of the proposed domain of study. Therefore, we provide the review of current systems developed using the social media data processing using ML based techniques.

Detecting fake news is a crucial problem. Existing techniques of fake news detection are based on supervised learning, which takes significant time and effort to build an effective and reliable dataset. ***S. Yang et al [1]*** investigate unsupervised manner for detecting fake news. Authors consider the true news and users' credibility as variables, and utilize users' engagements to identify opinions towards the authentic news. A Bayesian network is used to capture the dependencies among the true news, users' opinions, and the users' credibility. Then propose an efficient collapsed Gibbs sampling approach to classify truths of news and the users'

credibility. Experiment on two datasets show that the method outperforms.

The social networking sites are a target for the spammers to spread huge irrelevant information. For example, twitter is one of the most used platforms that allow a large amount of spam. Mostly fake users send undesired tweets for promotions that affect legitimate users and also consume resources. The expanding invalid data through fake users has increased results in the harmful content. The **detection of spammers and identification of fake users** has become a common research area. ***F. Masood et al [2]*** perform a review of techniques used in spammer detection. Moreover, taxonomy of the spam detection is presented. They classifies the techniques based on: (i) fake content, (ii) spam URLs, (iii) spam trending topics, (iv) fake users. The techniques are compared based on features, such as user, content, graph, structure, and time.

Y. Long et al [3] proposes a method to incorporate speaker profiles into an attention based LSTM model for **fake news detection**. The profiles used in two ways, first is to include in the attention model. Second is including profile as additional input. By adding profiles attributes the model outperforms the state-of-the-art method by 14.5% in accuracy. This proves that profiles provide valuable information to validate the credibility of news.

The Fake profiles are a type of malicious users, who is involved in various cyber crimes. Hence, their detection and removal become necessary to keep safe legitimate users and maintain trust. This system describes an approach to **classify fake vs. real profiles** based on various features. ***P. Shahane et al [4]*** use supervised learning classifiers i.e. Random Forest and Deep Convolutional Neural Network (CNN) and tested on Twitter dataset.

The most of human information are gathered and misused from online social media using fake profiles. ***V. A. Pashwan et al [5]*** has make efforts to ***detect the fake user*** on TWITTER, using support vector machine (SVM). This method identifies the fake user using profile attributes, with the accuracy of 97.33%. For future a work they suggest to extend the model by considering other attributes and use of various algorithms for more accuracy.

With the rapid growth of social networks, many problems like fake profiles and impersonation have also grown. ***S. D. P. Reddy et al [6]*** provide a framework for automatic ***identification of fake profiles*** efficiently. This model uses Random Forest Classifier to classify the profiles. It can be applied easily by online social networks that have millions of profiles.

Instagram has widely used for sharing photos and videos and is profitable for celebrities, businesses, and people with a number of followers. This high profit made this platform prone to be used for malicious activities such as ***fake accounts***. ***S. Sheikhi et al [7]*** provide a method for identifying Instagram fake accounts. First, a dataset of legitimate and fake accounts is created. Then, the dataset has been used with bagging classifier to classify fake users. The method compared to the five ML classifiers in terms of accuracy. The results show that the method performs better than other considered algorithms and correctly classified 98% of the fake accounts.

To prevent violating the policies of social media and also to avoid detection by a system like Google's Conversation AI, racists have begun to use a code to **distribute hate**. Users have used the words Google and Bing to represent the African-American and Asian communities. By generating the list of users who post such content, ***R. Magu et al [8]*** move a step forward from classifying tweets by allowing us to study the usage pattern of these concentrated set of users.

N. Shah et al [9] demonstrate a solution to address the challenges of **real-time analysis** using Elastic search engine and using distributed database. The database contains pre-build indexing and standardizing the search framework for large scale text.

The rapid growth of online health social websites has captured a vast amount of information and made easy to access. E-patients use these websites for informational and emotional support. ***L. Jiang et al [10]*** investigates the approaches for measuring user similarity. By similar users, we can help them to seek informational and support. They propose to represent the healthcare social media data as a heterogeneous healthcare information network and introduce the local and global structural approaches for measuring user similarity. Authors compare the approaches with the content-based approach. Experiments on a dataset from a health website showed that content-based approach performed better for inactive users, while structural approaches better for active users. Moreover, global structural approach outperformed for all users. In addition, local and global structural approaches using different weight schemas demonstrated that hybrid method yielded better than the individual approach. The global approach can deal with sparse networks and capture the implicit similarity.

In regular discussions, spellings, grammar and sentence structure are usually neglected. This may prompt various sorts of ambiguities to analyze and extract data patterns. ***S. A. Salloum et al [11]*** aims at analyzing textual data from Facebook and attempts to find knowledge and represent it in different forms. Findings indicated that Fox news is the most shared posts, followed by CNN and ABC News. The most frequent linked words are focused on USA elections. Moreover, the people are interested in sharing the news of Mohammed Ali Clay.

A. Singh et al [12] proposes a big-data analytics approach for Twitter to identify supply chain management issues. That includes text analysis using a SVM and hierarchical clustering. The result included a cluster of words which inform decision makers about customer feedback and issues. A case study in the beef supply chain was analyzed from Twitter data.

M. Alkhatib et al [13] propose a framework for ***events and incidents' management*** in smart cities. It uses text mining,

text classification, named entity recognition, and stemming to extract the intelligence. The data is automatically collected then processed to generate incident intelligence reports and provide situational awareness and early warning to rescue teams. The framework was implemented and tested using datasets from Arabic Twitter, and the results show that Polynomial Networks and SVM are the top performers, achieving accuracy of 96.49% and 94.58%.

It is important to analyze and classify data which lead companies to manage their works better. Social media may be containing various types of unwanted and maleficent spammer or hackers. ***Z. R. Mohi et al [14]*** choose Apriori for mining and classifying social media data and take Facebook for case study then after classifying data applying RSA algorithm to implement secure data.

The described review shows that the twitter is the most popular data source for experimental use. Then, we found some of the authors are also utilizing the Facebook, Instagram and other social media data. Similarly, in order to analyze the

Table 1 Review summary

Ref No	Domain	Algorithms	Dataset	Contribution
[1]	Detecting fake news	Bayesian network, collapsed Gibbs sampling approach	LIAR, BuzzFeed News	investigate unsupervised manner for detecting fake news
[2]	Detection of spammers and identification of fake users	NA	NA	A review of techniques used in spammer detection, classifies the techniques based on: (i) fake content, (ii) spam URLs, (iii) spam trending topics, (iv) fake users.
[3]	Fake news detection	LSTM		incorporate speaker profiles into an attention based model in two ways, adding profiles attributes model outperforms and achieve 14.5% higher accuracy
[4]	Classify fake vs. real profiles	Random Forest and Deep Convolutional Neural Network (CNN)	Twitter dataset	detection and removal become necessary to keep safe legitimate users and maintain trust
[5]	detect the fake user	SVM	TWITTER	This method identifies the fake user using profile attributes, with the accuracy of 97.33%.
[6]	identification of fake profiles	Random Forest Classifier	Online social networks that have millions of profiles.	classify the profiles
[7]	fake accounts	bagging classifier	Instagram	Correctly classified 98% of the fake accounts, a dataset of legitimate and fake accounts is created.
[8]	Hate speech detection	usage pattern, ML classifier	Twitter	generating the list of users who post such content
[9]	Real-time analysis in elastic search	pre-build indexing and standardizing the search	distributed database	address the challenges of using Elastic search engine, framework for large scale text
[10]	Accessing information made easy	user similarity, local and global structural approaches for similarity	online health social websites	A heterogeneous healthcare information network, the local and global structural approaches for measuring user similarity. And compare the approaches with the content-based approach.
[11]	spellings, grammar and sentence structure may prompt various sorts of ambiguities to analyze and extract data	NA	Facebook	Analyzing textual data, find knowledge and represent it in different forms.
[12]	chain management issues	big-data analytics, SVM and hierarchical clustering	Twitter data	Inform decision makers about customer feedback and issues. A case study in the beef supply chain
[13]	events and incidents' management in smart cities	Text mining, classification, named entity recognition, and stemming. Polynomial Networks and SVM.	datasets from Arabic Twitter	The data is automatically collected then processed to generate incident intelligence reports. Provide situational awareness and early warning to rescue teams. Achieving accuracy of 96.49% and 94.58%
[14]	analyze and classify data of companies to manage works better	Apriori, RSA	Facebook	Social media data for case study then after classifying data applying RSA algorithm to implement secure data.

data using ML algorithms different algorithms are applied. Among them SVM, CNN, LSTM, Bayesian Network and random forest is the main classifiers. Finally we found that the classification and analyzing the social media can be helpful in various social and business prospects. Additionally, the social media data can be utilized in many ways for human welfare. Among various applications and research orientations such as, fake news, fake users, spam detection, security, emergency management, accident management, awareness programs and many more directions are feasible with the social media. The next section demonstrate a basic ML model for social media data analysis.

III. PROPOSED WORK

A. Support vector machine (SVM)

A support-vector machine (SVM) is a binary classifier which finds a hyperplane or set of hyper-planes to distinguish the given patterns. The SVM algorithm can be used for classification, regression, and outlier detection. A classification performance is depends on hyperplane selection. For better classification accuracy we need to get hyper plan with largest distance to the training-samples. The aim of classification is to maximize the margin to minimize the error. For instance, in a finite-dimensional space when the sets to classes are not linearly separable. Then original dimensions will be mapped into a higher-dimensional space for easier separation. To keep computation reasonable, the mappings are designed to ensure that dot products of input vectors may be easy, by defining them in a kernel function $k(x, y)$ according to the problem.

The hyper-planes are defined as the set of points whose dot product with a vector is constant. The vectors defining the hyper-planes can be chosen to be linear combinations with parameters α_i of feature vectors x_i in the dataset. The points x in the feature space are mapped into the hyperplane are defined by the relation:

$$\sum_i \alpha_i k(x_i, x) = constent$$

If $k(x, y)$ is small as y grows away from x, each term in the sum the degree of closeness of the test point x corresponding dataset point x_i. The sum of kernels can be used to measure the relative closeness of test points.

B. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is data processing technique based on the concept of biological nervous systems. The ANN is structured as the network, which is composed by interconnected elements known as neurons. The configuration is depends on specific application, such as recognition, classification or prediction. Learning of ANN has involves update to the synaptic connections. The ANN is a complex, nonlinear, and parallel system. The basic unit of neural network is neuron, it consist of N no of inputs to the network are represented by $x(n)$ and each input are multiply by a connection weight which are represented by $w(n)$. The product of input and weight are summed and feed through a transfer (activation) function to generate the result (output).

IV. RESULTS ANALYSIS

V. CONCLUSION

REFERENCES

[1] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, H. Liu, “Unsupervised Fake News Detection on Social Media: A Generative Approach”, The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

[2] F. Masood, G. Ammad, A. Almogren, A. Abbas, H. A. Khattak, I. U. Din, M. Guizani, And M. Zuair, “Spammer Detection and Fake User Identification on Social Networks”, VOLUME 7, 2019, IEEE

[3] Y. Long, Q. Lu, R. Xiang, M. Li, C. R. Huang, “Fake News Detection Through Multi-Perspective Speaker Profiles” Proceedings of the 8th International Joint Conference on Natural Language Processing, pages 252–256, Taipei, Taiwan, November 27 – December 1, 2017,AFNL

[4] P. Shahane, D. Gore, “Detection of Fake Profiles on Twitter using Random Forest & Deep Convolutional Neural Network”, International Journal of Management, Technology And Engineering

[5] V. A. Pashwan, D. S. Ravi, “Fake Profile Detection on Twitter using SVM Classifier”, Indian Journal of Signal Processing (IJSP) ISSN: 2582-8320, Volume-1 Issue-1 February 2021

[6] S. D. P. Reddy, “Fake Profile Identification using Machine Learning”, International Research Journal of Engineering and Technology (IRJET), Volume: 06, Issue: 12, Dec 2019

[7] S. Sheikhi, “An Efficient Method for Detection of Fake Accounts on the Instagram Platform”, Revue d'Intelligence Artificielle, Vol. 34, No. 4, August, 2020, pp. 429-436

[8] R. Magu, K. Joshi, J. Luo, “Detecting the Hate Code on Social Media”, Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

[9] N. Shah, D. Willick, V. Mago, “A framework for social media data analytics using Elasticsearch and Kibana”, Springer Science+Business Media, LLC, Springer Nature 2018

[10] L. Jiang, C. C. Yang, “User recommendation in healthcare social media by assessing user similarity in heterogeneous network”, Artificial Intelligence in Medicine 81 (2017) 63–77

[11] S. A. Salloum, M. Al-Emran, K. Shaalan, “Mining Social Media Text: Extracting Knowledge from Facebook” International Journal of Computing and Digital Systems ISSN (2210-142X) Int. J. Com. Dig. Sys. 6, No.2 (Mar-2017)

[12] A. Singh, N. Shukla, N. Mishra, “Social media data analytics to improve supply chain management in food industries”, 2017 Elsevier Ltd All rights reserved

[13] M. Alkhatib, M. El Barachi, K. Shaalan, “An Arabic social media based framework for incidents and events monitoring in smart cities”, 2019 Elsevier Ltd All rights reserved

[14] Z. R. Mohi, “Apriori Method of Mining Secure Data in Social Media”, J Of University Of Anbar for Pure Science: Vol.13: No.1: 2019