International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 3 (2019), pp. 440-445
© International Research Publication House. http://www.irphouse.com

Sentiment Analysis on Social Media Data Using Intelligent Techniques

Kassinda Francisco Martins Panguila
PG Scholar, Department of Computer Science, CHRIST (Deemed to be University), India.

Dr. Chandra J.
Associate Professor, Department of Computer Science, CHRIST (Deemed to be University), India.

Abstract

Social media gives a simple method of communication technology for people to share their opinion, attraction and feeling. The aim of the paper is to extract various sentiment behaviour and will be used to make a strategic decision and also aids to categorize sentiment and affections of people as clear, contradictory or neutral. The data was preprocessed with the help of noise removal for removing the noise. The research work applied various techniques. After the noise removal, the popular classification methods were applied to extract the sentiment. The data were classified with the help of Multi-layer Perceptron (MLP), Convolutional Neural Networks (CNN). These two classification results were checked against the others classified such as Support Vector Machine (SVM), Random Forest, Decision tree, Naïve Bayes, etc., based on the sentiment classification from twitter data and consumer affairs website. The proposed work found that Multi-layer Perceptron and Convolutional Neural Networks performs better than another Machine Learning Classifier.

Keywords: Convolutional Neural Networks (CNN), Emotions, Machine Learning, Multi-layer Perceptron (MLP), Sentiment Analysis.

I. INTRODUCTION

With a big evolution of social media, the web enhances the appearance of a vibrant and lively realm in which billions of individuals around the world interact, share, post, and conduct numerous daily activities. Social media enables people to be linked together and interact with each other anywhere and anytime [1]. Social Media provides a various method for many people to precise and gives the opinion on a current or past event and many other activities around us [2]. More than 500 million people in the world give their opinion and views daily on the web [3]. A large quantity of knowledge is generated from various and different social media in numerous formats, numerous languages in the world. It makes challenges in data analytics to search out the new purpose and extract information from it [4, 5].

The Social Media mining is not exclusively Machine Learning strategies or other intelligent methods to identify and extracts information for sentiment analysis. On other hands, it is essential to know and determine the various domain for which unrelated gathering knowledge from the different place within the word, different time zone, language and separate values to be analyzed from entirely different perspectives [4].

This work attempts to recognize and recover useful facts and acquire facts from social media data such as tweets from a famous person and consumer affair data of Uber ride reviews to understand emotions, requirements to satisfy people needs. This approach is achieved by collect real-time data, clean and preprocess using various methods, and finally, identify the polarity by using binary classification and application of different intelligent techniques, and Neural Network approaches.

II. RELATED WORK

Various product, company, movies reviews have been studied in the field of sentiment analysis. To analyse political talks on social media [6] methodology was proposed to get more wise insights from data. Machine Learning approach was used to implement an approach for describing political abuse on Twitter [7] with the combination of topological and content-based feature extraction methods 96% of accuracy obtained with twitter data of US midterm elections in 2010. During the 2012 US election [8], sentiment analysis model for real-time data was proposed with the primary objective to identify real or fake political events and was tested with the various domain.

Sent meter an application for analysing opinion data was used during a campaign of Swachh Bharat Abhiyan in India in 2014 using an unstructured data from Twitter, and it achieves 84.47% of accuracy using machine learning approach with unigram words [9].

With the development of spatial-based Bayesian model [10], with an objective to find political leaders, a method applied for social media data for the US and some European countries. Various data from social media used during the legislative election of Spain in 2010 and US election [11] by analysing the user behaviour and political representatives. It shows that most people are from the urban area.

Lexicon based sentiment analyzer was proposed to determine the polarity and measures for tweet data of particular candidate [12]. A novel for unsupervised and supervised methods to determine the detection for sentiment analysis [13] with a supervised model they achieve 84% of accuracy that was better than Baseline approach but it takes much time to gives the result.

An unsupervised semantic arranged neural system approach utilizing Ant grouping calculation has been proposed to classify the sentences against a domain-specific tree obtained 44.38% as the best accuracy result [14].

The provided method to use dependency information with a derived vector from reliance tree using lexicon and SVM was generated with an accuracy of 70.2% using sentiment lexicon approaches [15].

Naïve Bayes and Support Vector Machine (SVM) are the most used machine learning approaches to formulate the sentiment analysis classifying into positive, negative or neutral. Nowadays researchers are using unsupervised learning using natural language processes and other intelligent techniques such as Neural Network approach to improve the accuracy but working better with a huge amount of data.

III. METHODOLOGY

The present section describes the methodology and various machine learning classifier used in this study. During real-time and sentiment analysis on social media data is difficult to know and say that a particular technique will be useful to complete the task, figure 1 represents the proposed methodology.

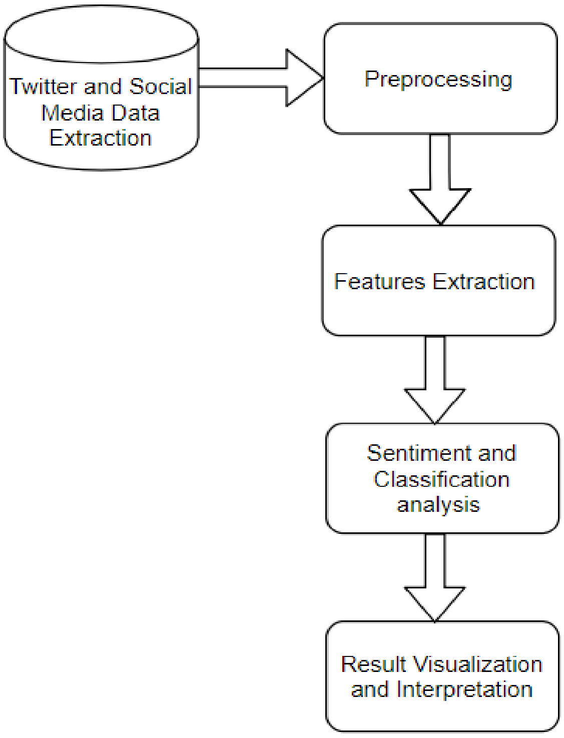
Twitter and Social Media Data Extraction
Preprocessing
Features Extraction
Sentiment and Classification analysis
Result Visualization and Interpretation

Figure 1. Proposed System

A. *Data Extraction*

The datasets is in the form of comma-separated values, one with tweets that were extracted at real-time twitter API connecting it to TAGS that is a google sheet template which lets us set up and ran an automated collection of data searched from Twitter [16], unique tweets of famous person across the world were collected, with the search term to get the result from the past seven days. The other dataset related to uber ride reviews that were extracted at real-time using beautiful soup's python library from a consumer affair website [17].

B. *Preprocessing*

Raw review scratched from the internet usually results in a noisy dataset. This is due to the casual nature of people's usage of social media. A vast number of preprocessing were applied to normalize the dataset and minimize its size.

C. Feature Representation

The ambiguity in word meaning is one of the major challenges in sentiment analysis. A word may be positive in some situation or be harmful in others [18]. Features extraction is critical to identify sentiment polarity.

TABLE I. SUMMARY OF UNIQUE UNIGRAM AND BIGRAM FEATURES EXTRACTED

Datasets	Unigrams	Bigrams
Uber Ride	7801	62288
Famous Personality	6813	21508

The unigram and bigram features were extracted from the datasets, and also the frequency distribution was created severally, Table I represent the number of unique unigram and bigram extracted from the datasets. Probably unigram as the most commonly used features for text classification is the presence of single words or tokens in the text. Bigrams are word pairs in the data which occurs in succession in the corpus and these features are an excellent way to model negation in Natural Language.

1) Spare Vector Representation

Nowadays a Spare Vector method has been more popular, where is applied in various fields with benefit [19]. One example is sentiment analysis in social media data, where depending on whether or not bigram features are used, the additional vector has a positive value at the indices. The positive value at the indices of unigrams and bigrams depends on the feature type specified which is one of presence and frequency.

- Presence: In the case of presence feature type, the feature vector has a 1 at indices of unigrams and bigrams present in the tweet and 0 elsewhere.
- Frequency: In the case of frequency feature type, the feature vector has a positive integer at indices of unigrams and bigrams which is the frequency of that unigram or bigram in the tweet and 0 elsewhere. A matrix of such designation frequency vectors is built for whole training dataset and then each term frequency is scaled by opposite file frequency of designation (*idf*) to assign higher values to important terms. The opposite file frequency of designation *k* is defined as:

$$idf(k) = \log\left(\frac{1+n_d}{1+df(d,t)}\right)+1 \tag{1}$$

Where, n_d is the absolute number of file and $df(d,t)$is the number of files in which the designation *k* occurs.

2) *Dense Vector Representation*

Various machine learning algorithms need the input to be represented as fixed-length of feature vector [20]. For a dense vector representation, an integer index was assigned to each word, which means that most common word is assigned the number 1, the second number 2 and so on. Each survey is then expressed by the vector of these lists which is a dense vector.

III.I. Classifier

In Machine Learning, classification is a directed learning approach in which the computer program gains from the information input given to it and after that utilizes this figuring out how to group new perception.

1) *Convolutional Neural Networks (CNN)*

Convolutional Neural Networks or CNN is a sort of neural systems which includes layers called convolutional layers which can translate spatial information. A convolution layer has a few channels or portions which it figures out how to remove the explicit kind of feature from the information. The Kernel is in the form of a 2D window which is slid at the end of information playing out the convolution activity. Temporal convolution was utilized in the current experiment which is appropriate for breaking down consecutive information like tweets or any other social media data.

2) *Multi-layer perceptron (MLP)*

MLP or Multi-Layer Perceptron is a class of feed-forward neural networks, which has at least three layers of neurons. Each neuron uses a non-linear activation function and learns with supervision using the backpropagation algorithm. It performs well in complex classification problems such as sentiment analysis by learning non-linear models.

3) *Naïve Bayes*

Naïve Bayes is a simple model which can be used for text classification. Mainly based on probability theorem that is known for making extraordinary performing yet clear models, particularly in the fields of document classification and disease prediction [21]. In this model, the class ĉ is assigned to a row r where:

$$\hat{C} = argmax_c\, P(c|r)$$

$$P(c|r) \propto P(c) \prod_{i=1}^{n} P(f_i|c) \tag{2}$$

In the formula above, f_i represents the *i*-th feature of total n features. P(c) and $P(f_i|c)$can be obtained through maximum likelihood estimates.

4) *Maximum Entropy*

Maximum Entropy classifier model is focused on the principle of Maximum Entropy. It is the leading general methods to find probability distribution from data and has been used for many natural language tasks [22]. The idea behind it is to choose the most uniform probabilities model that maximizes the entropy with given constraints. Unlike Naïve Bayes, it doesn't expect that features are restrictively independent of one another. In a binary classification problem like this approach, it is equivalent to utilizing Logistic Regression to discover circulation over the classes, and the model is represented by:

$$P_{nd}(c|d,\lambda) = \frac{exp\left[\sum_i \lambda i f_{i(c,d)}\right]}{\sum_{c'} exp\left[\sum_i \lambda i f_{i(c,d)}\right]} \tag{3}$$

Here, *c* is the class, *d* is the row and λ is the weight vector. The weight vector is found by numerical improvement of the lambdas to augment the contingent on probability.

5) *Decision Tree*

The Decision tree is the classifier model in which every node of the tree appears as a test on the feature of the dataset, and its progeny symbolize the endings [23]. The leaf node represents the last classes of the information. It is a supervised classifier model which utilizes information with realized names to shape the decision tree and after that, the model is connected on the test data.

6) Random Forest

Random forest is an ensemble learning algorithm for classification and regression. Random forest generates a multitude of decision trees classifies based on the aggregated decision of those trees. It is the most popular ensemble technique of classification because of the presence of best features [24]. For a set of rows $x_1, x_2, \dots x_n$ and their respective sentiment labels$y_1, y_2, \dots y_n$ bagging repeatedly selects a random sample (X_b, Y_b) with replacement. Each classification treef_i is trained using a different random sample (X_b, Y_b) where *b* ranges from 1…*B*. Finally, a majority vote is taken of predictions of these *B* trees.

7) Support Vector Machine (SVM)

SVM (support vector machines) is a non-probabilistic double straight classifier. For a training set of points $(x_i, y_i)$where$x$ is the feature vector, and$y$ is the class, with a need to find the maximum margin hyperplane that divides the points with$y_i = 1$ and$y_i = -1$. The equation of the hyperplane is the following:

$$w \cdot x - b = 0 \tag{4}$$

Maximizing the margin, denoted by, $max_{w,\gamma}\,\gamma, s.t. \forall i, \gamma \leq y_i(w \cdot x_i + b)$ in order to separate the points well.

IV. RESULT AND DISCUSSION

A. Sentiment and emotion Analysis

Uber as one of the famous company of a distributed ridesharing, taxi, bike sharing and a transportation network organization is important to understand expectations, emotions and opinion of people. A political famous person like Donald Trump and others denoted as one of the most popular politic leaders in the world is needed to understand the emotions, requirements of people. This can be achieved using intelligent techniques to analyze the data.

TABLE II. EMOTION COUNTS

Datasets	Emotions	Counts
Uber Ride Review	Negative	523
	Positive	433
	Neutral	388
Famous Personality	Negative	394
	Positive	508
	Neutral	332

Sentiment analysis is done expecting to determine the demeanor of people for a particular subject or relevant disparity of a document. The straightforward assignment in sentiment analysis is to categorize the polarity of a sentence or manuscript as positive, negative or neutral as described on Table II with respective counts from the documents where figure 2 and 3 shows the graphical representation of with respective percentage found.

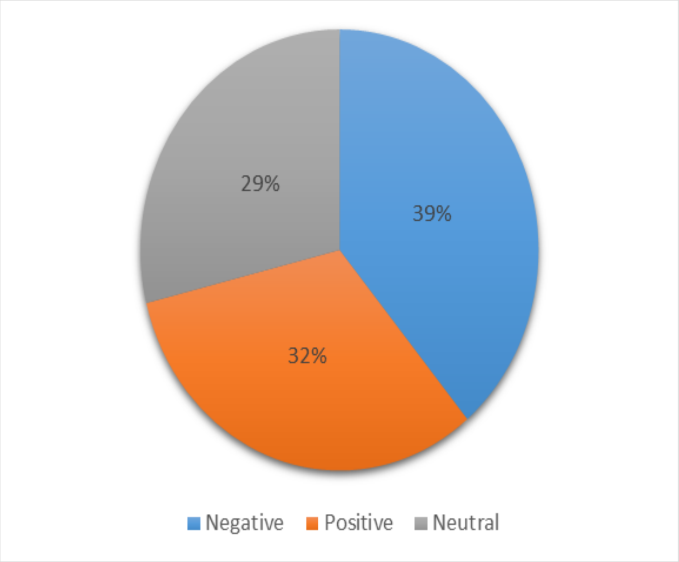

Figure 2. Sentiment Analysis of Uber Ride Review

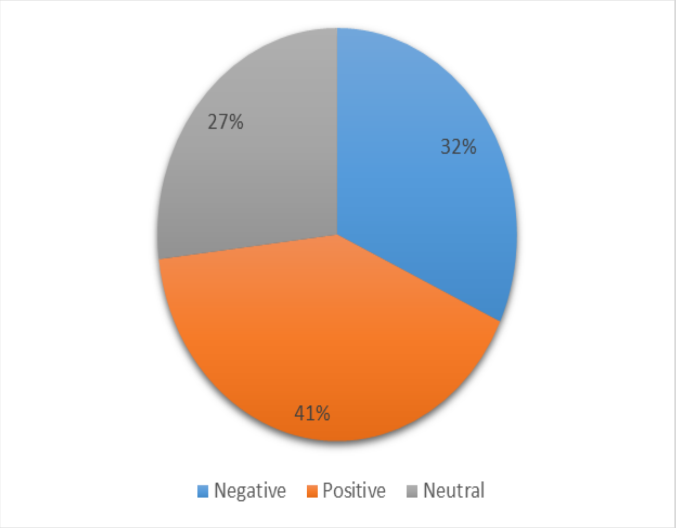

Figure 3. Sentiment Analysis of Famous Person

B. Experiments

The experiments were performed on various classifiers. Unless otherwise specified, 10% of the training datasets were used for validation of models to check against overfitting. A Sparse vector representation of several reviews and tweets were used for Naïve Bayes, Maximum Entropy, Decision Tree, Random Forest, SVM and Multi-layer Perceptron (MLP) the respective accuracies is described in Table III. The dense vector representation of each review was used to implement the Convolutional Neural Networks (CNN) that gives better results.

TABLE III. COMPARISON OF VARIOUS CLASSIFIER

Algorithms	Uber Ride Review	Famous Person
Naïve Bayes	83.74%	74.19%
MaxEnt	87.41%	80.02%
Logistic Regression	93.40%	91.41%
Decision Tree	81.48%	79.40%
Random Forest	84.44%	78.19%
SVM	83.70%	84.33%
MLP	96.71%	95.9%
CNN	96.85%	97.01%

Figure 4 and figure 5 shows the graphical representation of performance evaluation on the various classifier, where the neural methods as described above gives better accuracy with related to the present social media data provided.

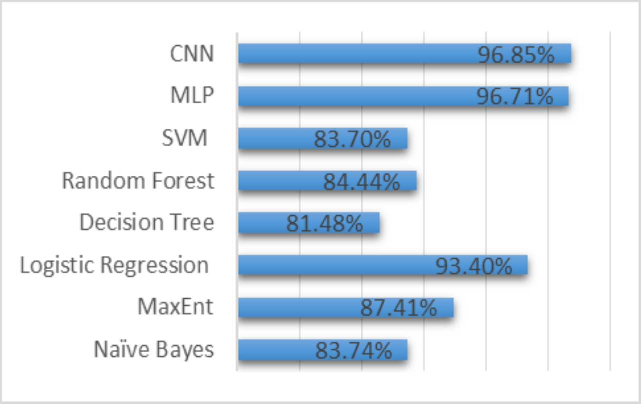

Figure 4. Performance Evaluation on Uber Ride Review

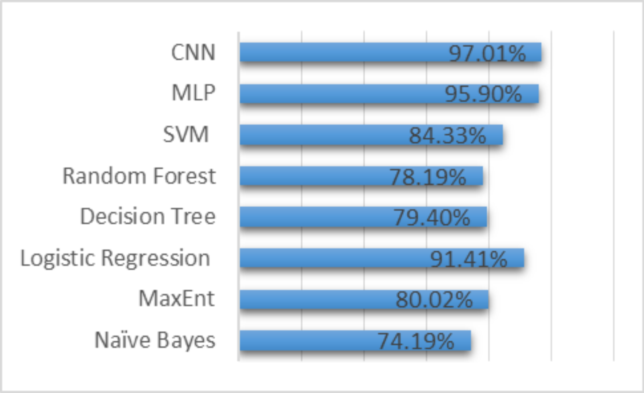

Figure 5. Performance Evaluation on Famous Person

V. CONCLUSION

Sentiment analysis using intelligent techniques approach in this paper was proposed to deal with social media data. It has been observed that various techniques can be used to achieve a sentiment analysis on social media data and others. However, with the methods used shows that presence in the spare vector representation recorded a better performance than frequency. According to the experiment result of Twitter data and uber ride data from consumer affair website shows that Neural Networks methods such as Multi-layer Perceptron (MLP) and Convolutional Neural Network (CNN) performed better than others classifier in general. Whereas, its proposed system can be applied in the other internet community.

REFERENCES

[1] P Reza Zafarani, Mohammad Ali Abbasi, Huan Liu, Social Media Mining – An Introduction, Cambridge University Press, Publisher Location, 2014.

[2] Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S.Li, L. T., Shoemaker, D. J., …Xie, L. Social media use by the government: from the routine to the critical. Government Information Quarterly, 29(4), 2012.

[3] Twitter Usage Statistics – Internet Live Stats. (n.d.). Retrieved October 22, 2018, from http://www.internetlivesstats.com/twitter-statistics.

[4] Liu, B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 2012.

[5] Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, Social media analytics. Business and Information system Engineering (2014), 89-996.

[6] Stieglitz, S., & Dang-Xuan, L., Social media and political communication: a social media analytics framework. Social Network Analysis and Mining 3(4), 2013, 1277-129.

[7] Ratkiewiez, J., Conover, M., Meiss, M., & Gonçalves, B, Detecting and Tracking Political Abuse in Social Media. ICWSM, 2011.

[8] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. A System for Real Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. Proceedings of 50th Annual Meetings of the Association for Computational Linguistics, 2.12, 115-120.

[9] Tayal, D. K & Yadav. Sentiment Analysis on Social Campaign "Swachh Bharat Abhiyan" using unigram method. AI & Society, 2016.

[10] Barbera, P. Birds of the Same Feather tweet together. Bayesian ideal point estimation using twitter data. Political Analysis, 2015.

[11] Barbera, P., & Rivero, G. Understanding the Political Representativeness of Twitter Users. Social Science Computer Review (2015), 712-729.

[12] Farha Nausheen, Sayyada Begum, Sentiment analysis to predict election results using python, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Explore Compliant, 2018, 978-1-5386-0807-4.

[13] K. Schouten, O.V. Weijde, F. Frasincar, and R. Dekker, Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-Occurrence Data, IEEE Transactions on Cybernetics, 2017.

[14] E. S. Chifu, T. Leția, and V. R. Chifu, Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps, In Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on, pp. 1-9. IEEE, 2015.

[15] K. Paramesha, and K. C. Ravishankar, Exploiting dependency relations for sentence level sentiment classification using SVM, In Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on, pp. 1-4, IEEE, 2015.

[16] Tags Google Sheet Template, Version 6.0, from http://tags.hawksey.info/.

[17] Consumer Affairs, https://consumeraffairs.com /travel/uber.html.

[18] Sneha Passerate, Rajashree Shedge, Comparative Study on Feature Extraction Techniques used Sentiment Analysis, international conference on innovation and challenges in cyber security (ICICCS), 2016.

[19] Y. Yuan, X. Li, Y. Pang, X. Lu, and D. Tao, Binary sparse nonnegative matrix factorization, IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 5, pp. 772–779, 2009.

[20] Quoc Lec, Tomas Mikolov, Distributed Representations of Sentences and Documents, Google Inc.

[21] Sebastian Roschka, Naïve Bayes and Text Classification, Corner University Library, 2014.

[22] Kamal Nigan, John Lafferty, Andrew McCallum, Using Maximum Entropy for Text Classification, IJCAI workshop on Machine Learning.

[23] Qing-Yun Dai, Chun-Ping Zahang and Hao Wu, Research on Decision Tree Classification Algorithm in Data Mining, International Journal of Database Theory and Application vol.9, 2016.

[24] Eesha Goel, Er. Abhilasha, Random Forest: A review, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), 2017.