International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN (P): 2249–6890; ISSN (E): 2249–8001
Vol. 10, Issue 3, Jun 2020, 6127-6134

TRANS STELLAR
•Journal Publications •Research Consultancy

Original Article

BIG DATA APPROACH OF SENTIMENT ANALYSIS OF TWITTER DATA USING K-MEAN CLUSTERING APPROACH

MUDASSIR KHAN[1], AADARSH MALVIYA[2] & SURYA KANT YADAV[3]

[1]*Research Scholar, Noida International University, Greater Noida, India*

[2]*Assistant Professor, Noida International University, Greater Noida, India*

[3]*Assistant Professor & Head, Noida International University, Greater Noida, India*

ABSTRACT

The social networking sites are storming with vast amount of information is produced and processed. This production of enormous amount of data from different microblogging platforms needs to be stored and processed by Big data approach. The data generated in peta bytes from social media platforms is called big data. Most of the world populations are sharing their opinion and feedback every day on micro blogging sites, the opinion and behavior of the user on microblogging site is tedious. Number of devices, laptops, iPad, Tabs, MacBook and distinct IOT data gadgets produced enormous amount of data. The processing and retrieving of this huge amount of data becomes a very complicated task. Nowadays social media is playing a crucial rule in our day to day life. The social media platforms have been greatly used for expressing human opinions for the item and verities of services. The business of different organizations is based on the feedbacks and rating of millions of the microblogging site uses can be used to extract their opinion, attitudes and sentiment towards the products. This feedback information on social media used for the future market of the business improvement and analysis of the product performance. The extracting the user's opinion from the social media platforms is a difficult task; it can be redefined into various ways. In this paper, an open source approach is presented, and data is collected from tweets from twitter API, analyzed and visualized these tweets using python. To analyze sentiments of tweets we have used K-Means clustering algorithm and implemented with Python programming language. This sentiment analysis is based on the tweets data retrieval from different sources and then classifying the user perspectives in different classifications accuracy and two unique sentiments (positive and negative).The process of sentiment analysis is the computational procedure that deciding the output by the proposed method is positive or negative. In this paper, an endeavor has been modelled to prefer analysis method for sentiment of twitter dataset. In the proposed technique polarity of each tweet is calculated to differentiate whether the tweet is neutral, positive or negative. The sentiment sentence's polarity is the emotions of user such as angry, sad, happy and joy. The proposed entire research work processed and implemented using Python.

KEYWORDS: K-means, Clustering, Sentiment analysis, Bigdata, Bigdata sentiment & Twitter data

Received: Jun 10, 2020; **Accepted**: Jun 25, 2020; **Published**: Aug 03, 2020; **PaperId**.: IJMPERDJUN2020580

1. INTRODUCTION

The social networking sites are used by millions of people around the world every day to demonstrate their opinions, attitudes, and behavioral feelings towards their business into a short message. The short messages can get the popularity on the social media platform within a short period of time within a connected community. Nowadays the microblogging platforms are gaining extreme popularity among individuals, celebrities, organizations and business communities to share their opinion and feedback with their followers. The huge number of messages are posted every second; this has been become the main idea behind the success of any business venture to gain the

popularity through the social media platforms. The collective sentiments toward business would be very helpful in different ways, e.g., an organization can use the feedbacks of the products and further enhance its products or can redesign marketing campaigns to achieve the customer targets, the government organizations can analyze the people's rate of satisfaction can resolve their problem on time. The twitter is the most popular microblogging platform used by trillions of users which allows its users to post a tweet within 140 characters. The tweets contain the user's annotations, abbreviations, links, images, hashtags and emotions. The sentiment analysis of these features of users is a challenging task till now. There are different types of machine learning techniques are available to handle sentiment analysis smoothly, but clustering is the unsupervised machine learning approaches that needed no labels while practicing forming a group of similar information into a cluster.

This type of cluster can be utilized to basic sentiments concerning an event or entity. The K-means is a robust and extremely fast approach. The K-means locates K centroids, every centroid enlist the closer data into its cluster by using Euclidean distance an rectify its centroids by the mean of all the data present within the entire cluster. The operation is repetitive until the centroids have no coinage. The short tweet or communication can cause the major sparseness and with extremely high dimensional dataset that build K-means to drop its cost-effectiveness when trading with this type of dataset.

The K-means also dependable on the inceptive state of centroids which regularly connect to an existing local optimum. To overcome these drawbacks, the proposed method will be described in the following sections. After solving the drawbacks, the next approach is to find the optimal number of clusters K, that gives best quality of data description. After completing the clustering process, each message within a cluster will be pointed by screening the positive and negative terms using Sent WorldNet.

2. RELATED WORKS

The main objective of sentiment analysis is to extract the subjective data that proceeds author's opinion toward something or anyone by using existing NPL (natural language processing), text analytics, computational linguistics, etc., that is based on documentation-level or sentence-level classification. Most of the researchers thinks that sentiment analysis of microblogging messages is similar as sentiment analysis at sentence level [2,3]. The polarity if words and phrases used in the sentences to write messages on the social media sites is a very challenging task. The different attributes of the microblogging messages using distinct slangs, abbreviations, links, hashtags, and generated emotions that may execute well at sentence or document level but might not work in microblogging platforms.

The main objective of the project was to find a solution to differentiate between fake and genuine banknotes using the K-means clustering algorithm, which separates data into K-clusters. Forging of banknotes has been increasing day by day and thus many banks have faced big problems trying to identify those banknotes. This report can do that prediction and identification which hopefully save time and cost considerably

3. PROPOSED METHOD

The approached used in this research is categorized in distinct main stages: Pre-processing stage, clustering stage, and interpreting stage. The Pre-processing stage used on different social media platforms, Twitter, was mainly used for the study case. Twitter message (tweets) can be expressed with a maximum of 140 characters. A Twitter user who has been introduced in the tweet will be deployed by an @ symbol. A hashtag symbol is used to help the other Twitter users easily search the trending tweets and they can contribute their valuable input in the trending tweet by a # symbol.

The data or rather the dataset which was given to create the project was names “Twitter authentication dataset”. In the dataset there are two feature- V1 and V2.

To work on this research the mean and standard deviation of both V1 and V2 was calculated. Where the mean of V1 and V2 is 0 .434 and 1.922 respectively. And the standard deviation of V1 and V2 are 2.842 and 5.867 respectively

As mentioned, the dataset in this report is consisted of data extracted from imagines with Wavelet Transform. The imagines were taken from genuine and forged banknote specimens (n=1372). There are two attributes in this dataset (V1: variance of Wavelet Transformed image and V2 skewness of Wavelet Transformed image).

Table 1 summarizes the two variables. It shows that V2 has higher mean and standard deviation and wider range than V1. Figure 1 visualizes the distribution of the data.

Table 1: V1 V2 Mean 0.433735 1.922353 Standard Deviation 2.842763 5.869047 Min -7.042100 -13.773100 Max 6.824800 12.951600

		Table 1	
	V1		V2
Mean	0.433735	1.922353	
Standard Deviation	2.842763	5.869047	
Min	-7.042100	-13.773100	
Max	6.824800	12.951600	

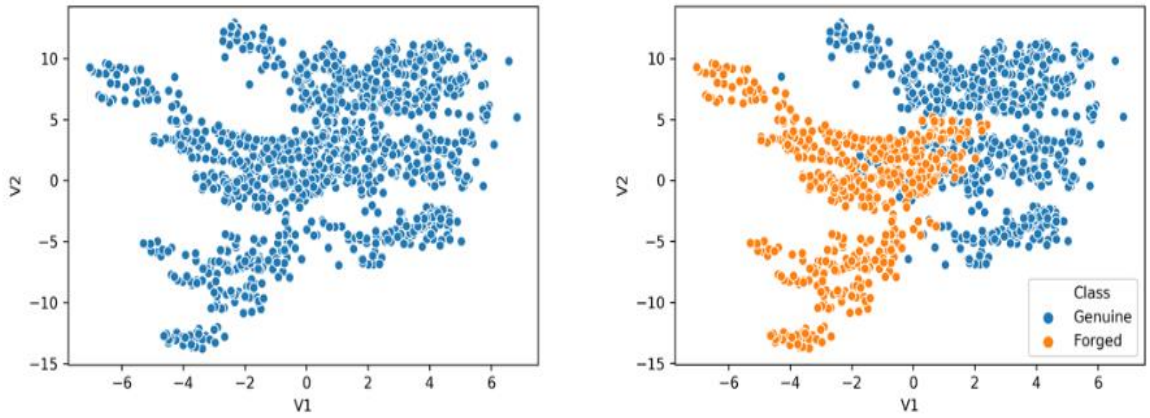

Figure 1: (Left) The Distribution of Data without Labels (Right) The Distribution of data with Labels

3.1 Clustering Stage

K-mean cluster analysis is suitable to analyze this dataset since (1) there is enough datapoint in the dataset (no missing values) (2) the numbers of instances are more than the numbers of the features.

In order to make the variable comparable, two variables were first normalized so both variables are in the range of 0 and 1. Then k-mean cluster analysis was conducted. It is aimed to make two cluster (i.e. forged and genuine). Figure 2 visualizes and color-codes the two clusters from the analysis. Two black dots are the centers of the two cluster.

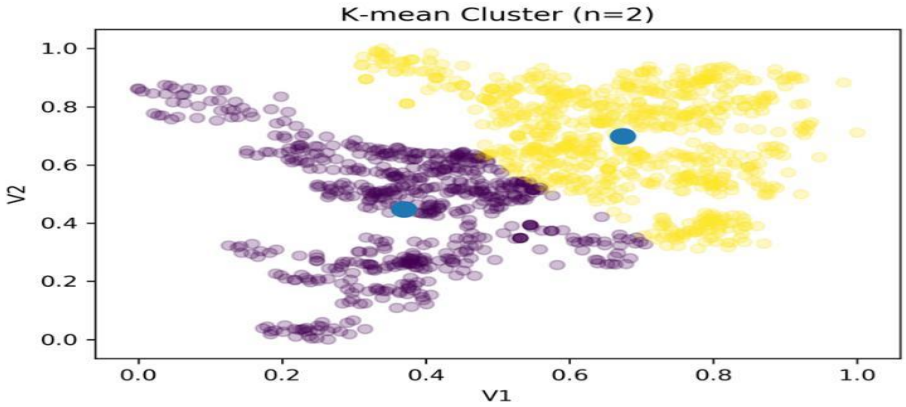

Figure 2: Visualization of K-Mean Cluster Analysis

From visualizing the twitter data, we have found the graph shown above. To evaluate if the dataset is suitable for K-Means, some sentiment analysis points were checked. First of all, the given data of the both features are of value float which is good for the algorithm. There are many missing data points generated by the k means clustering which do not provide any problem while implementing the algorithm. After visualizing the graph, we can have overview that there are some outliers which should not create enough problems for the existing data because most of the twitter datapoints are inside the ellipse as shown in the graph. Next, we have given enough training data (1372 instances) with many features. So, the numbers of extracted features are much less than the instances exist in the dataset which ensure the absence of curse of dimensionality. Lastly, K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. That's why the dataset was normalized.

3.1.1 Clustering using K-means and Interpretation

Figure 2 is the visual representation to compare the K-mean cluster results and the actual labels of the dataset. With coarse visual inspection, it looks like K-mean cluster can successfully distinguish features extracted from twitter dataset. However, it is also noted that there are some overlaps, and some are possibly misidentified. The sentiment analysis of the extracted featured is somehow forged.

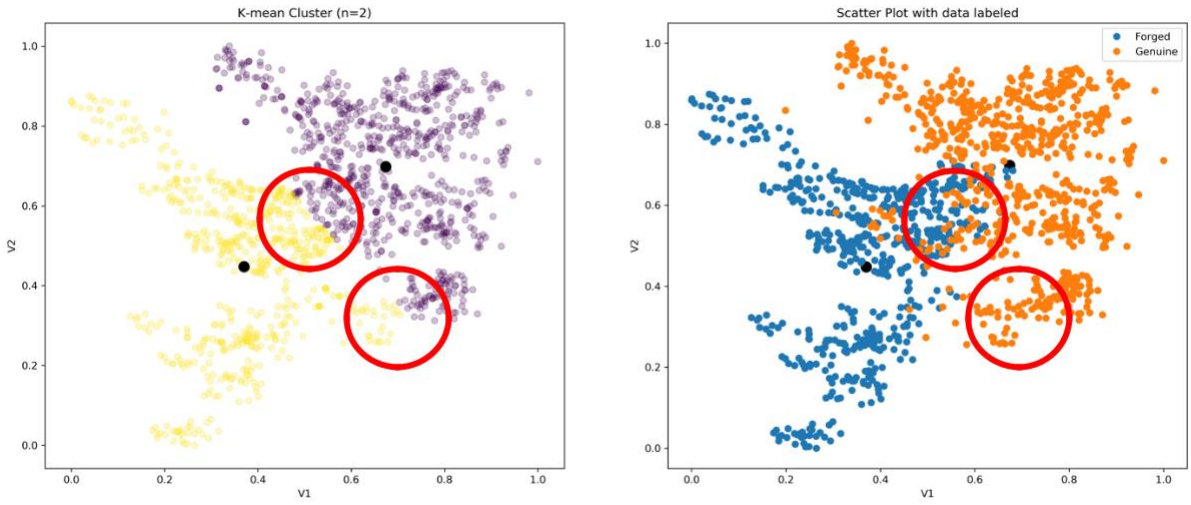

Figure 3: (Left) The Results of k-mean Cluster. (Right) The Scatterplot with Data Labeled

Table 2 turns the visual representation into numbers. Compared the prediction from the K-mean cluster analysis and the data label from the dataset. It shows that error rate is 12.76% with half of it false positive (i.e. misidentified forged tweets as genuine tweets) and half of it false negative (i.e. misidentified genuine tweets as forged tweets). The success rate in distinguishing between forged sentiment of tweets and genuine sentiment of tweets from the extracted features accuracy is 87.24%.

Table 2: Summary of the Sentiment Accuracy

	Number of Sentiment Accuracy		%
	Match	1197	87.24
Mismatch	False Positive	89	6.486
	False negative	86	6.268

Detecting sentiment from tweets has become a recurring problem in the microblogging websites these days. Various researchers all over the world are opting this field of big data for analysis of sentiment. Security measurements are taken places, such as ultraviolet and holographic features. It is advised that a Wavelet Transform is the most cutting-edge tool right now. Hence, B bank recently obtained a dataset that contained data extracted from images that were taken from genuine and forged banknote specimens using a Wavelet Transform tool (n=1372). The ultimate aim is to use this dataset to train a machine to detect fake notes automatically.

However, before implementation, it is important to access if this dataset can sufficiently distinguish forged banknote from genuine ones. Hence, in this report, with k-mean cluster analysis, unsupervised machine learning, performed on the dataset, we will visualize and outline the results and make according recommendation.

Sentiment analysis of Twitter data based on big data approach has become a trending field for the researchers all over the world. The existing and new popular social network platforms are generating huge amount of data that is called big data, Twitter had been the point to interest to thousands of researchers in important areas like prediction of democratic events, customer brands, movie box-office, stock market, reputation of personalities etc. The big data approach of sentiment analysis refers to the feelings or opinion of person towards some particular domain or an event. The extracting or making an experimental approach for analysis of sentiment (opinions) and its classification based on polarity of sentences, document or a short message is a challenging task. The different approaches of the challenges are overwhelming amounts of data on one particular topic with all having distint representation. The Classification and clustering are two majors widely used methods applied to perform sentiment analysis of twitter data and big data. We have used K-Means clustering algorithm to find sentiment performance classification that improves accuracy on tweets.

4. CONCLUSIONS

The success rate in distinguishing between forged sentiment of tweets and genuine sentiment of tweets from the extracted features accuracy is 87.24%. The sentiment classification performance of the twitter data gives better result in terms of classification accuracy 87.24%.The algorithm is working stable. Having two clusters, there visibility of cluster around the 0 variance and 0 skewness of the banknote images, although more towards the negative quadrant of the graph, while the other cluster appears at high skewness and higher variance coordinates. in this context, it could suggest that real twitter data will be represented by the first cluster around the origin of the scatter plot, and the fake around the second cluster in the higher skewness and variance coordinates. The three-cluster scenario further confirms a cluster around the origin.

5. RECOMMENDATIONS

The research paper is to be used as solution to real life problems of huge amount data that is called big data, but the model of the research can be much more efficient if more features were used and if more data was given. And also, the model may not be able to differentiate always as new better sentiment analysis of twitter data are created day by day to make the model obsolete hence the model must be given new data every once in a while, to make sure the model is up to date. Overall, the current dataset with K-mean cluster analysis can successfully detect most of the sentiment analysis from twitter dataset from the genuine datasets, with the successful rate of 87.24%. As long as the bigdata is content with the success rate of accuracy, this dataset and k-mean algorithm have proven to be effective in distinguishing the classification accuracy and performance of sentiment analysis. The dataset and the methods are recommended. If the sentiment analysis process of big data is wised to increase the classification accuracy, more data will be needed to avoid false positive.

REFERENCES

1. *S. Baccianella, A. Esuli and F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in Proc. Int. Conf. Language Resources and Evaluation (2010), pp. 2200–2204. 2. H. Yu and V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in Proc. 2003 Conf. Empirical Methods in Natural Language Processing (2003).*
2. *S.-M. Kim and E. Hovy, Determining the sentiment of opinions, in Proc. 20th Int. Conf. Computational Linguistics (2004), pp. 1367–1373.*
3. *E. Kouloumpis, T. Wilson and J. Moore, Twitter sentiment analysis: The good the bad and the OMG!, in Proc. Fifth Int. AAAI Conf. Weblogs and Social Media (AAAI Press, 2011), pp. 538–541.*
4. *A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proc. Int. Conf. Language Resources and Evaluation (2010), pp. 1320–1326.*
5. *A. Agarwal, B. Xie, I. Vaovsha, O. Rambow and R. Passonneau, Sentiment analysis of Twitter data, in Proc. Workshop Language in Social Media (2011), pp. 30–38.*
6. *B. Xiang and L. Zhou, Improving Twitter sentiment analysis with topic-based mixture modeling and semi-supervised training, in Proc. 52nd Annu. Meeting Association for Computational Linguistics, Vol. 2 (2014), pp. 434–439.*
7. *J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2011) 1–8.*
8. *W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri and C. Haruechaiyasak, Discovering consumer insight from Twitter via sentiment analysis, J. Univ. Comput. Sci. 8 (2012) 973–992.*
9. *P. Paulraj and A. Neelamegam, Improving the performance of K-means clustering for high dimensional data set, Int. J. Comput. Sci. Eng. 3 (2011) 2317–2322.*
10. *K. Krishna and M. N. Murty, Genetic K-means algorithm, IEEE Trans. Syst. Man Cybern. 29 (1999) 433–439.*
11. *D. Karaboga and An idea based on honey bee swarm for numerical optimization (Technical Report No. TR06, Erciyes University, 2005).*
12. *D. Karaboga and C. Ozturk, A novel clustering approach: arti¯cial bee colony (ABC) algorithm, J. Appl. Soft. Comput. 11 (2011) 652–657.*
13. *G. Armano and M. R. Farmani, Clustering analysis with combination of arti¯cial bee colony algorithm and k-means technique, Int. J. Comput. Theory Eng. 6 (2014) 141–145.*

14. *C. Musto, G. Semeraro and M. Polignano, A comparison of lexicon-based approaches for sentiment analysis of microblog, CEUR Workshop Proc. 1314 (2014) 59–68. SA on Microblogging with K-Means and ABC 1950017-21 Int. J. Comp. Intel. Appl. 2019.18. Downloaded from www.worldscientific.com by 5.82.170.170 on 06/25/20. Re-use and distribution is strictly not permitted, except for Open Access articles.*

15. *G. A. Miller, WordNet: A lexical database for english, J. Commun. ACM 38 (1995) 39–41.*

16. *A. Hamzehei, M. Ebrahimi, E. Sha¯ee, R. K. Wong and F. Chen, Scalable sentiment analysis for microblogs based on semantic scoring, in Proc. 2015 IEEE Int. Conf. Services Computing (2015), pp. 271–278.*

17. *M. Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in Proc. 5th Annu. Int. Conf. Systems Documentation (1986), pp. 24–26.*

18. *S. Banerjee and T. Pedersen, An adapted lesk algorithm for word sense disambiguation using WordNet, in CICLing 2002: Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol. 2276 (Springer, Berlin, 2002), pp. 136–145.*

19. *Khan, M. and M. D. Ansari (2020). "Multi-criteria software quality model selection based on divergence measure and score function." Journal of Intelligent & Fuzzy Systems 38(3): 3179-3188.*

20. *Khan, M. and A. Malviya (2020). Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.201, 978-1-7281-4142-8/$31.00 ©2020 IEEE.*