International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195429

Data Mining Approach of Text Classification and Clustering of Twitter Data for Business Analytics

P. Meghanasree[1], Dr. S. Gopi Krishna[2]

[1]PG Scholar, Department of Computer Science and Engineering, Sri Mittapalli College of Engineering, U9, Thummalapalem, Prathipadu, Guntur, Andhra Pradesh, India

[2]Professor, Department of Computer Science and Engineering, Sri Mittapalli College of Engineering, U9, Thummalapalem, Prathipadu, Guntur, Andhra Pradesh, India

ABSTRACT

The increasing popularity of micro-blogging sites like Twitter, which facilitates users to exchange short messages (tweets) is an impetus for data analytics tasks for business development. Twitter has a huge amount of data. Twitter's API allows you to do complex queries like pulling every tweet about a certain topic. So, Companies can know more about consumers' sentiments towards their products and services and use them to better understand the market and improve their brand. In this paper selected a popular food brand to evaluate a given stream of customer comments on Twitter. Several metrics in classification and clustering of data were used for analysis. A Twitter API is used to collect twitter corpus and feed it to a classifier algorithm that will discover the polarity lexicon of English tweets, whether positive or negative. A clustering technique is used to group together similar words in tweets in order to discover certain business value.

Keywords : Sentiment Analysis, Classification, Clustering, Twitter, Data Mining

I. INTRODUCTION

The social media has reanalyzed the nature of how companies approach their business development processes. The social media contains a huge volume of unstructured data like tweets and reviews. This paper of analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and they can be formally defined as the process of finding out "useful" patterns in a big volume of data set, either automatically (unsupervised) or semi-automatically (supervised). The paper would heavily depend on with full trust or confidence on techniques of NLP ("Natural Language Processing") in extracting significant patterns and characteristics from the large data set of tweets and on "Machine Learning" techniques for precisely classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them. The characteristics that can be used for modeling patterns and classification based on natural language. Hence, the industry mainly focuses on social media research [1]. In fact, [2] forecasted that by 2024, the market for text analytics will rise to $11.29 billion with a growth rate of 18.2%. The report sed that how companies depend on social media for social media data analytics and predict feature of the company and the pattern of their business.

This paper mainly catechizes the use of sentiment analysis in business development. Furthermore, this paper SUBSTANTIATE the text analysis process in reviewing the public opinion of customers towards a certain product that can be used for making a decision after the text analysis is completed. More so, stressed that there is limited academic literature Related to text analytics of Twitter data, as a result, this paper focus on developing an application for mine the tweeter data and clustering customer tweets.

II. METHODS AND MATERIAL

Twitter is a social and micro-blogging service that allows users to post real-time experiences as messages, it's called tweets. Tweeter allows short messages, restricted to 140 characters in length. Due to the character of this micro-blogging service (quick and short messages), people use some special characters, make wrong words, use some different icons and other special characters that express proper meanings. Following is a brief terminology related to tweets. icons: These are persons facial expressions like smile and sad, etc pictorially represented using punctuation and letters; and they express the user's mindset.

Sentiment Analysis is one of the processes for analyzing users' sentiments and their opinions regarding company products and services. Companies can utilize this feature for developing their business. Sentiment analysis is a process that gets the users reviews or tweets like users opinions, views, emotions, tweets, etc. this data can be saved in the database and convert into datasets(CSV). Analyzing this data by using NLTK(one of the NLP package).

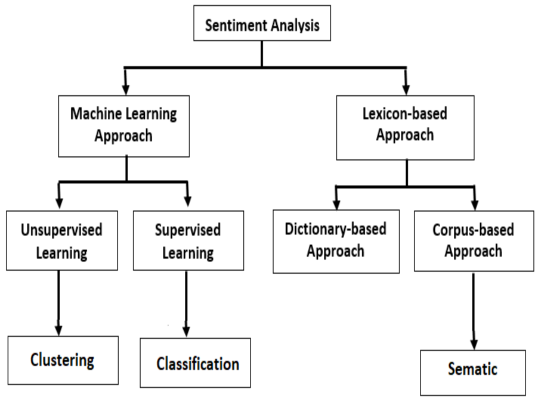

Fig. 1 sentiment analysis

Several types of research have raised out the importance of customer sentiment analysis for business operations. Majority of the situations this is helpful for business improvement and decision-making purpose. Sentiment analysis is with data mining was proven to be most efficient to find quality and quantity for customer satisfaction. This is more helpful for companies to increase the growth of product development and profits.

In this experiment we acquire 10,000 manually annotated Twitter data (tweets) from a tweeter commercial source by the using of tweeter API. this data can be processed and arrange the proper format (data set). Now find the sentiment (positive, negative, natural) using Text Blob next we perform validations.

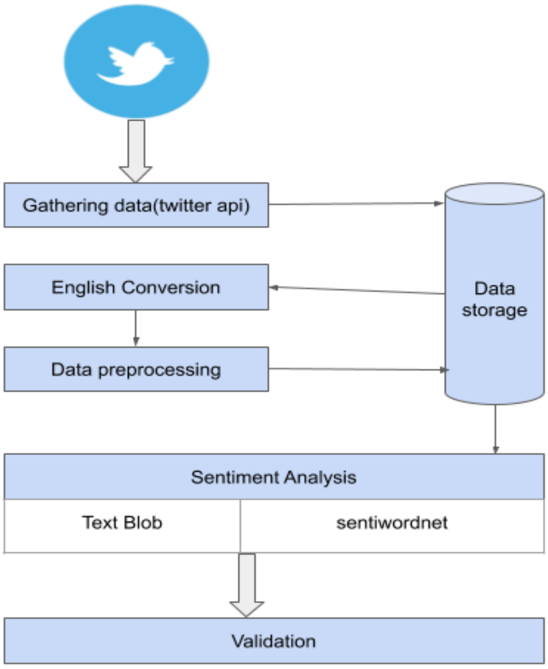

Fig. 2 tweet classification and sentiment analysis

III. PROPOSED METHODS

Sentiment analysis can provide valuable insights from social media platforms by detecting emotions or opinions from a large volume of data present in an unstructured format. Sentiment analysis includes three polarity classes, which are negative, neutral and positive. The polarity of each tweet is determined by assigning a score from –1 to 1 based on the words.

Negative values => Negative sentiment
Positive values => Positive sentiment
Zero value => Natural sentiment

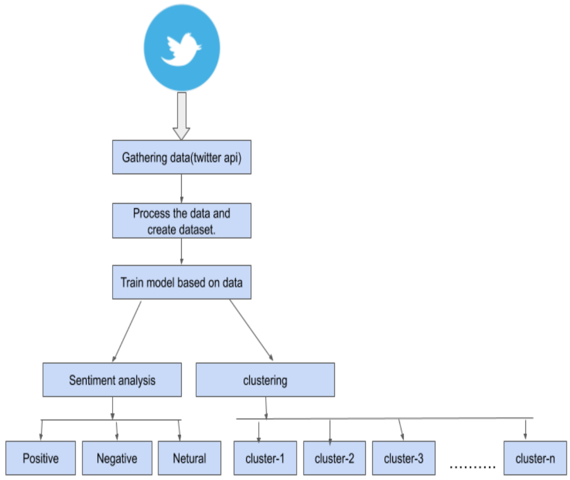

Fig. 3 tweets clustering and sentiment analysis

In order to validate the results obtained from TextBlob, we used analysis of the data. When we calculated the polarity and subjectivity of the Twitter dataset, the resultant files were in comma separated value (CSV) format. Figure 1b illustrates the model construction and evaluation process. For model building, we applied unsupervised machine-learning algorithms, k-means on the training dataset, and calculate accuracy for LSTM. Steps of the analyzers' validation through k-mean and LSTM validation can be viewed from the pseudocode "k-mean clustering" and " LSTM deep learning". These machine-learning algorithms (k-mean and LSTM) were applied to the training set to build an analysis model. On the basis of the model constructed for each analyzer, the test set was evaluated. After the test set evaluation, we recorded the accuracy of each analyzer under each model.

Clustering is an unsupervised machine learning method in analyzing the context of text data in natural language. It is a mathematical approach in collecting and segmenting similar words into clusters. It helps trim down the volume of unstructured text and provide a structured text. It also provides the keywords in each cluster that is useful in extracting valuable insights **Clustering** is unsupervised machine learning method in analyzing the context of text data in natural language. It is a mathematical approach in collecting and segmenting similar words into clusters. It helps trim down the volume of unstructured text and provide a structured text. It also provides the keywords in each cluster that is useful in extracting valuable insights, hence, customer sentiments can be summarized using these keywords. The clustering process is illustrated below fig.

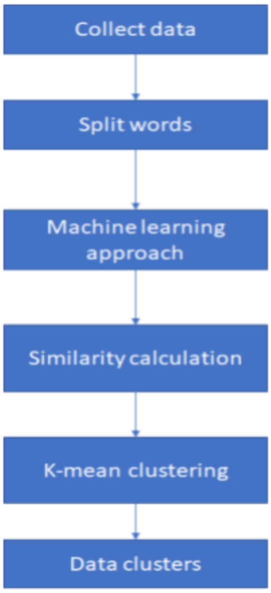

Fig.4 data clustering

K-means algorithm is a clustering algorithm used in cluster analysis that finds a user-specified number of clusters that are represented by their centroids. clustering is unsupervised learning, it does not need a label data. K-means is efficient to multiple runs. This research performed 10 repeated run times. It simply decides the set of k clusters and assigns each word into the cluster.

LSTM (Long Short Term Memory networks): LSTM network is a kind of RNN(recurrent neural network).

This is useful to estimate stock prices, electricity demand and so on. This is performed by spread back the output of a neural network layer at time t to the input of the same network layer at time t + 1. This LSTM is used for calculating the accuracy of the clustering approach for the final result.

IV. CONCLUSION

The consequences of evidence-based decision making contribute to the improvement of a brand. Having a text analysis of customer feedback and reviews allows for effective quality management. With sentiment analysis, companies can now plan to reposition their businesses according to customer's sentiments. This paper provided an introduction and reasoning behind the value of text analytics of Twitter data to businesses in gaining customer views on products and services and brand.

This paper also discussed several related works in sentiment analysis for business applications. Importantly, it demonstrated a practical application of text classification and clustering of Twitter data, and show ways on how to analyze these to gain business insights. Although the clustering accuracy rate for this experiment is already acceptable in this application domain. It is suggested that future work needs to increase the accuracy of the clustering model by improving data preparation and experimenting with other Clustering algorithms.

Future work in this field can also be focused on real-time analytics of Twitter data stream. Since there is a massive number of tweets collected daily, handling real-time analytics is difficult. Therefore, an automated sentiment analysis, which runs in high processing and large memory computing resources, is required.

V. REFERENCES

[1]. B. Liu, "Sentiment Analysis and Opinion Mining," Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1–167, 2012.

[2]. Marketsandmarkets.com, "Text Analytics Market by Component (Software, Services), Application (Customer Experience Management, Marketing Management, Governance, Risk and Compliance Management), Deployment Model, Organization Size, Industry Vertical, Region - Global Forecast to 20," 2017.

[3]. F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, "SentiBench - a benchmark comparison of state-of-thepractice sentiment analysis methods," EPJ Data Sci., vol. 5, no. 1, 2016.

[4]. A. Moreno and T. Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence," Int. J. Interact. Multimed. Artif. Intell., vol. 3, no. 6, p. 57, 2016.

[5]. V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," Int. J. Comput. Appl., vol. 139, no. 11, pp. 975–8887, 2016.

[6]. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093– 1113, 2014.

[7]. L. Ziora, "The sentiment analysis as a tool of business analytics in contemporary organizations," Stud. Ekon., pp. 234–241, 2016.

[8]. S. Yaram, "Machine learning algorithms for document clustering and fraud detection," in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017.

[9]. N. Yussupova, M. Boyko, and D. Bogdanova, "A Decision Support Approach based on Sentiment Analysis Combined with Data Mining for

Customer Satisfaction Research," Int. J. Adv. Intell. Syst., vol. 1&2, 2015.

ABOUT AUTHORS :

P. Meghanasree is currently pursuing her M.Tech (CSE) in Computer Science Department, Sri Mittapalli college of Engineering, Guntur, A.P. She receiving her M.Tech in CSE from SMCE, Tummalapalem.

Dr. S. Gopi Krishna is currently working as an Professor in Computer Science Department, Sri Mittapalli college of Engineering, Guntur, A.P. His research includes data mining and Machine Learning

Cite this article as :

P. Meghanasree, Dr. S. Gopi Krishna, "Data Mining Approach of Text Classification and Clustering of Twitter Data for Business Analytics", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 138-142, July-August 2019. Available at doi : https://doi.org/10.32628/CSEIT195429
Journal URL : http://ijsrcseit.com/CSEIT195429