Research Ambition ISSN: 2456-0146
An International Multidisciplinary e-Journal
(Peer - reviewed, Open Accessed & Indexed)
Web: www.researchambition.com Email: publish2017@gmail.com
Vol. 4, Issue-I
May 2019
Impact Factor : 5.936 (SJIF) ******** e-ISJN: A4372-3068

Application of Text Classification and Clustering of Twitter Data for Business Analytics

Sandhya Singh
Scholar of Computer Science
Bhabha Institute of Technology
cs.sandhya17@gmail.com

Anamika Tiwari
Faculty of Computer Science
Bhabha Institute of Technology
anamika1107.tiwari@gmail.com

Abstract –

In the ongoing years, interpersonal organization in business are increasing tremendous prevalence since it helps in the business development. Organizations become more acquainted with about their customers feelings of their item in market which encourages them to comprehend the market better and as needs be, they make their market and business technique. There is a colossal information accessible on interpersonal organization site like Twitter, Facebook and so on. Utilizing Machine Learning devices and strategies this information can be prepared into applicable informational collection so as to decide the example and patterns to increase significant bits of knowledge. This paper chosen a famous nourishment brand to assess a given stream of client remarks on Twitter. A few measurements were utilized to arrange and bunch information which was utilized for investigation. A Twitter API is utilized to gather information and channel it into a Binary tree classifier which will find the extremity of the tweets, regardless of whether positive or negative. A k-implies bunching strategy is utilized to aggregate comparable words so an important esteem can be given to the business. This paper endeavours to examine the specialized and business point of view of content order of Twitter information and features the future open doors this rising field.

Keywords – *Twitter, Sentiment Analysis, Decision Tree, k-means, Social media.*

I. INTRODUCTION

Social media has redefined the nature of how companies strategize their business processes. The social media contains huge volume of unstructured data in form of tweets, blogs, user post and reviews. All these data can be used for business intelligence such as consumer profiling and content analysis. Twitter, which is a social networking online platform is used to great extent for marketing and promotion tools by many companies. Specifically, twitter data not only contains user information, but also text that contains subjective information (such as user sentiments) towards a particular issue. From business point of view twitter data contains enough reviews of users which can be used by companies to get a feedback about their product or services. And all these a company can easily extract without spending a huge amount of money on customer surveys and interview. On the other hand analysing and extracting meaning full information form unstructured data poses a significant challenge for the data miners. Humans can easily find easily find the sentiment of the document but this is limited to small data set but when dealing with a huge corpus of data then some automated tooling and techniques are needed to deal with this problem. Hence the industry that deals with sentimental analysis is gaining momentum and currently the focus of social media research[1]. In fact[2] forecast that by 2022, the market for text analysis will rise to $8.99 billion with a growth rate of 17.2%. The report describes that increasing growth is due to increase in the demands of the data miners by the companies for their social media analytics, predictive analytics and customization of their business application.

This paper mainly examines the use if sentiment analysis in business application. Furthermore, the paper describes the text analysis process in reviewing the public opinion of the customer towards certain brand and presents hidden knowledge. so,[3] stressed that there is limited academic literature surrounding text analytics of Twitter data, as a result, this paper attempts to contribute in this developing field by providing a practical guide on how to mine and analyse customers' tweets. All these data can be used to make business decision after the data has been processed. This paper attempts to contribute in the developing filed by providing a practical guide on how to mine and analyse customer tweets.

This paper is divided into (4) sections. Following this section is Section II which provide background Information and related work of following topics: Sentiment Analysis, Decision Tree and Text Clustering. Section III demonstrate the Experimental setup combined with the result. Section (IV) presents the Conclusion and future work.

II. BACKGROUND AND RELATED WORK

A. Sentiment Analysis

Sentiment Analysis is used by companies to analyse corpus of data to understand user's sentiment or opinion regarding their product and services. According to [4] Sentiment Analyses is "a process that automates the mining of attitude, opinions, views and emotions from the text, speech, tweets and database sources through Natural Language Processing". Fig. 1 shows different approaches and techniques.

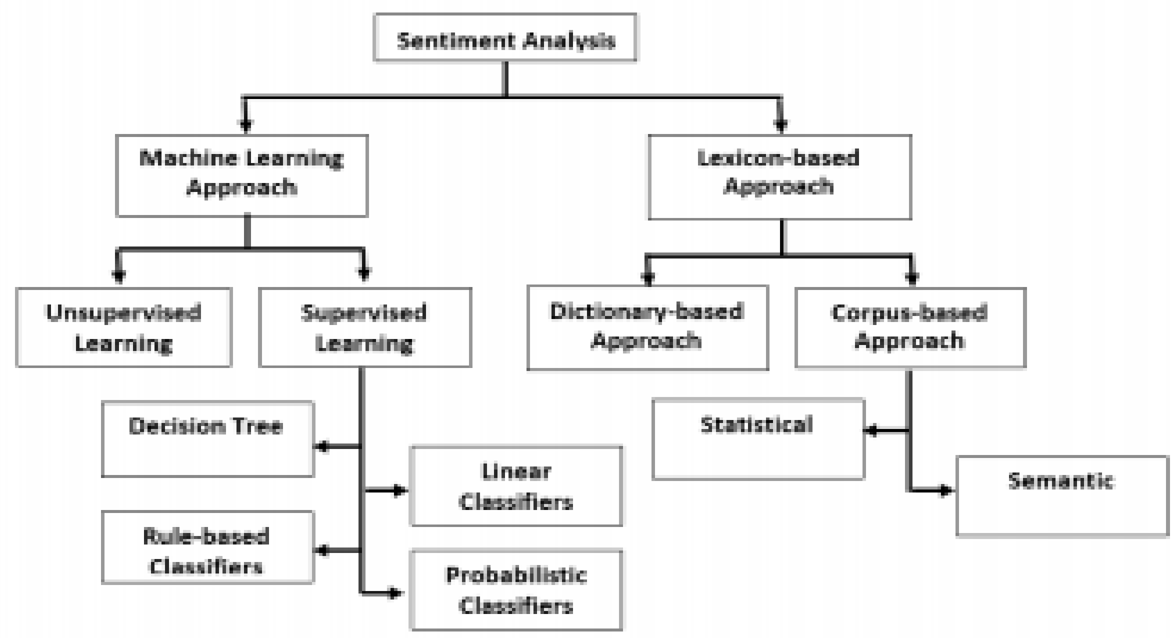

Fig. 1 Sentiment Analysis, source: [5, Fig 1]

Many studies have concluded the importance and benefits of customer sentiment analysis for business operations[6]. The study pointed out that customer satisfaction is the key factor that influences decision making, thus, it is important to know customers' sentiments to manage the quality.

B. Text Classification using a Decision Tree

Text classification is an automated process of classifying text in natural language to predefined categories[7]. It is best used for predicting binary or nominal class labels[8]. Algorithms involved in this are Naïve Bayes Classifier, Decision Trees, Support Vector Machines, Neural networks, and many others. The statistical approach in solving a classification problem involves two learning methods: supervised learning method (as shown in Fig. 2) which uses a training dataset to build the classification model prior to application in a test set, and unsupervised learning method which does not use any known labels in mining the data.

A decision tree, which is of interest in this paper, is an example of a supervised learning method. Decision tree is a machine learning technique that uses hierarchical data structure. Moreover Decision tree algorithms are used in data mining because they are proven to produce rational classification models and good accuracy levels.

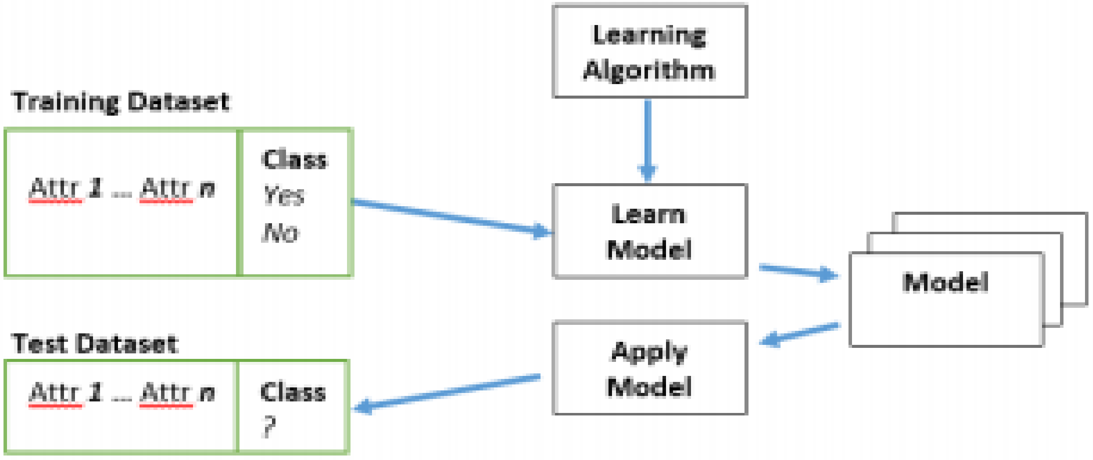

Fig 2. Text Classification building model, Source, [9, fig 4.3]

C. Text Clustering

Clustering is one of the commonly used unsupervised learning methods in analysing the context of text data in natural language form[10]. It is a mathematical approach in collecting and segmenting similar documents into clusters. It helps trim down the volume of unstructured text and provide a simpler understanding and thematic structure of the data.

K-means is a clustering technique in cluster analysis that finds a user-specified number of clusters that are represented by their centroids.

Basic k-mean algorithm is described as:

1. Select k points as initial centroids.
2. Repeat
3. Form k clusters by assigning each point to the closest centroid
4. Form k clusters by assigning each point to closest centroid
5. Recompute the centroid of each clusters
6. Until centroid do not change

The algorithm iteratively assigns each point to one of the k groups based on the specified features and the points are clustered based on feature similarity[11]. Firstly, the points are assigned to initial centroids. Each point is assigned to the nearest centroid and the centroid of that cluster is updated by computing the mean of all points that are assigned to the centroid' cluster. This process is repeated until a stop criterion is satisfied.

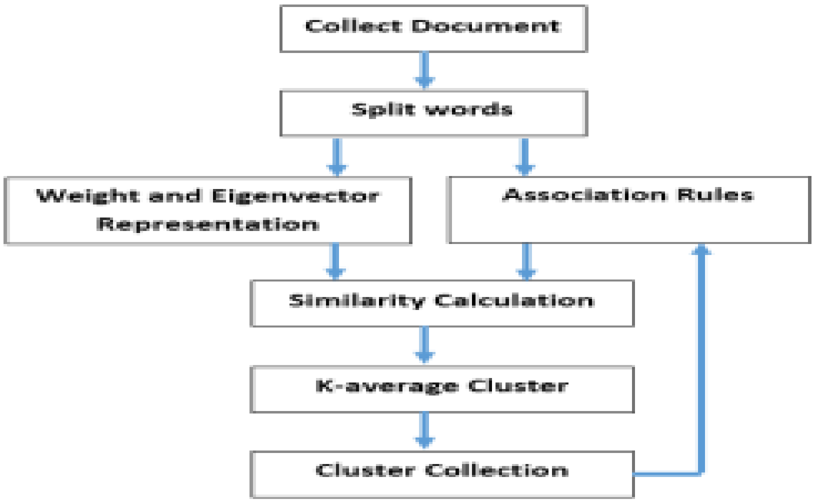

Fig. 3 text Clustering source [3, Fig 4]

III. EXPERINMENTAL SETUP

A. Software specification

Rapidminer is a data-driven data mining tool that discovers patterns in large collections of text[12]. It is an analytic platform that integrates machine learning and predictive model deployment used by data scientists. It contains rich libraries of data science and machine learning algorithms[13].

B. Hardware Specification

The experiment was developed using the following hardware specifications: Processor: Intel Celeron CPU N2830@2.16GHz 2.16GHz RAM: 4 GB (3.98 GB usable) System Type: 64-bit Operating System.

C. Text Classification Process Model

Tasks involved in the text classification are data extraction, data cleaning, pre-processing and feature extraction, and classification[14].

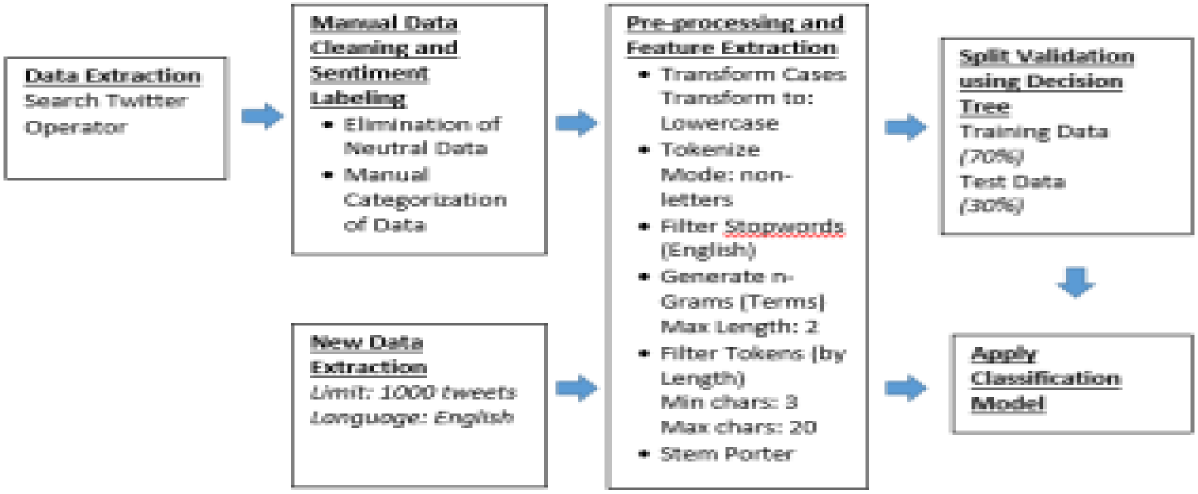
Data Extraction
Search Twitter Operator
Manual Data Cleaning and Sentiment Labeling
• Elimination of Neutral Data
• Manual Categorization of Data
New Data Extraction
Limit: 1000 tweets
Language: English
Pre-processing and Feature Extraction
• Transform Cases Transform to: Lowercase
• Tokenize Mode: non-letters
• Filter Stopwords (English)
• Generate n-Grams (Terms) Max Length: 2
• Filter Tokens (by Length) Min chars: 3 Max chars: 20
• Stem Porter
Split Validation using Decision Tree
Training Data (70%)
Test Data (30%)
Apply Classification Model

Fig. 4 Text classification Process in Rapidminer

Details of the process is shown in the diagram Fig 4. The process applied to this experiment are described in the succeeding sections.

1. Data Extraction

An access token is used as authentication mechanism to get the data from twitter database. This access token is generated by login to your twitter account. Authorisation token that include API key are required to establish the connection and allow a search query. In the search query parameter of the twitter API a popular food brand is used. Other search parameters include returned tweet and language limit of 500 English tweets, and the result type of recent or popular tweets. This contains attribute and label types including tweet ID, username, number of retweets, original text, date and time it was created, language, and many others. The dataset was saved in MS Excel file.

2. Manual Data cleaning and Sentimental Labelling

The saved data is set to be reviewed by human labeler whose task was to filter irrelevant data and retaining those which are significant to the experiment. The labeler was instructed to the label the text as the positive, negative or neutral with the following guidelines. Out of the 500 tweets, 352 neutral tweets were discarded since they serve no purpose in the analysis. The remaining dataset of 148 tweets were discarded since they serve no purpose in further experiment. The sentiment labeling is manually conducted by two (2) experts who classified each tweet as either positive or negative.

3. Feature Extraction

A pre-processing stage is needed in order to clean the dataset in order to train the model. Five operations were applied at the pre-processing stage.

They were as follow transform Cases, Tokenize, Filter Operator, Filter Token, Stem Porter to extract the useful feature of the data. The following operations are described as follow –

a) The Transform cases operator converts the characters in a document to lowercase so that the words, “Hello” and “hello” are the same.

b) The Tokenize operator splits the texts of a document into a sequence of words (or tokens) by setting the nonletters mode as splitting points.

c) The Filter Stop words operator removes the unwanted words from a document such as is, are, the, of, etc. These words are commonly used in the text but provide no content information.

d) The Generate n-Grams administrator makes nlength of tokens in a report. The administrator will check words that much of the time tail each other. In content examination, it is basic that tokens are gathered so as to separate all the more significance. Single tokens, for example, "wellbeing", "living", "summer", and "breeze" may give little data instead of combined (or 2 gram) tokens, for example, "healthy_living" and "summer_breeze". In this manner, it will be clearer the setting of the related terms.

e) The Filter Tokens (by Length) operator removes tokens based on their length. For instance, if the minimum and maximum characters are set to 3 and 20, respectively, then it will remove texts having less than 3 characters and more than 20 characters in length.

f) The Stem Porter administrator diminishes the length of the words until a base length is come to or when the words are in its base structure. For instance, the words "power" is abbreviated to electric", "restoration" is abbreviated to "reviv", "socialism" is abbreviated to "commun, etc. This activity makes the correlation between words simpler.

IV. CONCLUSION AND FUTURE WORK

The end result of evidence-based decision making contributes in the improvement of product brand. Hence effective quality management is achieved through analysing customer data using text analysis. So now companies can strategically reposition their businesses according to customers' sentiments.

This paper provides an introduction and reasoning behind the value of the text analysis of Twitter data to businesses in gaining customers views on product and services. Although the classification 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 classification model by improving data preparation and experimenting with other classification algorithms.

Future work in this field can likewise be centred around real-time investigation of Twitter information stream. Since there is a huge measure of tweets gathered day by day, taking care of continuous examination is troublesome. In this way a robotized feeling investigation, which keeps running in high handling and enormous memory figuring assets, is required.

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. 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.
4. 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.
5. 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.
6. L. Ziora, “The sentiment analysis as a tool of business analytics in contemporary organizations,” Stud. Ekon., pp. 234–241, 2016.
7. S. M. Kamruzzaman, F. Haider, and A. R. Hasan, “Text Classification using Data Mining,” Science (80-.)., p. 19, 2010.
8. T. Pang-Ning, M. Steinbach, and V. Kumar, Introduction to data mining. 2006.
9. O. Muller, I. Junglas, S. Debortoli, and J. Von Brocke, “Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data,” MIS Q. Exec., vol. 15, no. 4, pp. 64–73, 2016.
10. M. Allahyari et al., “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” arXiv Prepr. arXiv, vol. 1707, no. 2919, pp. 1–13, 2017.
11. A. Trevino, “Introduction to K-means Clustering,” Datascience.com. 2016.

12. M. Hofmann; and R. Klinkenberg;, "RapidMiner: Data Mining Use Cases and Business Analytics Applications," Zhurnal Eksp. i Teor. Fiz., 2013.
13. K. S. Rawat, "Comparative Analysis of Data Mining Techniques, Tools and Maching Learning Algorithms for Efficient Data Analytics," JOSR J. Comput. Eng., vol. 19, no. 4, pp. 56–60, 2017.
14. H. Kaur and V. Mangat, "Dictionary based Sentiment Analysis of Hinglish text," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 816–822, 2017.