INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Science, Technology & Management (ICSTM-2020)

Sentiment Analysis of Twitter Data

Sanjay Rai[1], S. B. Goyal[2],Jugnesh Kumar[3]
[1]Student,City University, MALASIA
[2]Director, Faculty of Information Technology,City University, MALASIA,
[3]Director, SAITM, INDIA
drsbgoyal@gmail.com[2]

Abstract

The World Wide Web has taken seriously new ways for individuals to convey their views and conclusions on different topics, models and issues. Clients create content that resides in a variety of media, such as web gathering, conversation gathering, and weblogs, and provide a solid and generous foundation for gaining momentum in different areas such as advertising and research. Policy, logic research, market forecasts and business outlook. Hypothesis research extracts inferences from information available online and orders the emotions that the author conveys for a particular item into up to three predefined categories (good, negative, and unbiased). Identify the problem. This article outlines a hypothesis review cycle for quickly ordering unstructured news on Twitter. In addition, we are exploring different ways to perform a detailed emotional survey on Twitter News. In addition, it presents a parametric correlation of strategies considered according to recognized boundaries. This work tends to make the case enjoy investigating on Twitter; The values communicated in them represent the tweets: positive, negative or fair. Twitter is an online thumbnail that contributes to a blog and a wide range of interactions, allowing customers to create short 140-character short instructions. It is a fast growing association with more than 200 million subscribers, of which 100 million are dynamic customers and half of them constantly sign up for Twitter, generating around 250 million tweets every day. Due to this overwhelming use, we plan to achieve a biased impression of the public by breaking the estimates communicated in the tweets. Researching public opinion is important for some applications, for example, when companies are looking to respond to their material, predict political careers, and anticipate economic wonders like stock trading. The function of this to build a useful classifier for the command in a precise and programmed way of the stream of fuzzy tweets.

Keywords: Twitter, Social media, Machine learning, Sentiment analysis, Natural language processing, Mining, Part of speech, Bigram

1. Introduction

Micro blogging websites have been evolving for becoming a source of data and information. This is due to the nature of microblogs, where individuals constantly post their feelings on a range of topics, investigating current issues and complaints, and sharing their guesses about everyday objects. The various number of people have been utilizing social media like Facebook, Instagram and Twitter. Online media can help you create large amounts of information such as tweets, status, remark, blog entries, surveys, and more. Different associations are looking for extracting data from twitter for analysing sentiments of the user over Twitter. This study will help counter the difficulties of apathetic investigation on Twitter, which may be associated with the characterization of tweets based on reliability, denial, and impartiality. The active users of Twitter has been huge all over the world. The test will focus on

researching customers based on their tweets, comments and messages. The main research center will help create a practical classifier to correctly estimate little-known tweet flow. Using the term "inquiry" will help confirm any subject matter of Twitter customer sentiment and the superiority provided by any demanding institution by the administration.

2. Background

3. As stated by Hassan et al., (2020), the use of microblogging has expanded lately with the focus of web-based and internet-based media. Social media has produced a huge amount of emotional information in terms of status, blogs and articles. It also allows businesses to connect with their customers through social media. Therefore, users have proven their opinions and comments about their services and products on social networks such as Twitter. Therefore, associations can analyze this data to understand the feelings of consumers. The use of microblogging platforms has improved in recent years. Previous years are seeing large enhancement in using of microblogging platforms. Sentiment analysis has been helping in understanding whether data related to products and services are satisfactory or not. As mentioned by Sailunaz&Alhajj (2019), textual data, extraction techniques have been helping in processing, analysing and searching facts from this data. Facts have been regarded as the objective component, however there are some subjective features. These contents have been related to opinions, appraisals, emotions and sentiments that help in conducting the base of sentiment analysis. This has caused many problems related to the development of new applications due to the high level of data. As commented by Kumar and Jaiswal (2020), sentiment analysis has been linked to a process that automates the mining of opinions, attitudes, and ideas from textual data using natural processing language. It incorporates the categorization of positive, negative and neutral opinions in the text. Some examples of sentiment analysis terminology are given below:

<SENTENCE> = the quality of the product was high

<OPINION HOLDER> = <author>

<OBJECT> = <product>

<FEATURE> = <quality>

<OPINION> = <high>

<POLARITY> = <positive>

Emotion analysis includes a variety of tasks, including emotion extraction, subjectivity classification, emotion classification, and opinion summary. It has been aiming at analysing sentiments of people, attitudes, emotion and opinions for elements involving products, topics, services and individuals (Nakov et al., 2019).

4. Related Work

As mentioned by Ruz, Henríquez&Mascareño , (2020), many researches and studies have been done based on sentiment analysis on reviews, opinions and news articles. Sentiment analysis focuses on the broader field of natural language processing, with an emphasis on the computer-assisted study of emotions, moods, and opinions expressed in text format. The analysis of opinions and feelings aimed to study people's opinions about attitudes and emotions towards the organization. The entity can be represented as individuals, events, and themes. The bigram model works well with sentiment analysis. As mentioned by Terán&Mancera , (2019), Part of speech and bigram model are not helping in outperforming with all models.`. The past years have been huge increment in utilization of microblogging platforms. Some of the popular microblogging websites including Twitter have been evolved for becoming a source of various types of information. Twitter was designed as a platform where users are accustomed to most of their reviews, feelings and thoughts. Data was diversified across platforms because it is based on real-time data. Therefore, this research will be based on Twitter as it is used by most people. As per view of Srivastava, Singh &Drall (2019), data available over the twitter is related to users who used to be professional on blogs. So, sentiment analysis can be performed on such data properly. A sense prescription system has been helpful in recommending the system with proper analysis. The use of this system helps to maintain a deep attitude to promote better analysis of emotions. This system will not recommend assets which used to receive lots of negative feedback and ratings. There have been various abusive language and negative words included in the collected data provided over the twitter. Therefore, these sentiments have been analysed for

understanding the better opportunity for association in the market. Negative sentiment has been helping companies in resolving issues with the products and services in the market (Hasan et al., 2018).

5. Literature Survey

As commented by Rodrigues &Chiplunkar (2019), sentiment analysis have been based on NLP working at different levels of granularity. This is monitored at sentence level and now at the phrase level. This research have been based examining sentiments of tweets that come under Pattern classification and data mining. The above terms are based on the process of inventing better patterns in data. This study will be based on techniques of NLP for mining patterns and features from large data sets. As per view of Saad & Yang (2019), machine learning is a technique used in differentiating individual tweets based on the pattern model. The features can be used for modelling patterns and distinguishing patterns and classification based on a couple of groups such as formal blogging and informal blogging. Language based characteristics deals with formal linguistics and include sentiment polarity of user thoughts. Sentiment polarity has been focused on words that are based in natural innate tendency. This study has been helping in addressing challenges based on sentiment analysis in Twitter that can be based on differentiating tweets on negative, positive and neutral. As per view of Srivastava, Singh &Drall (2019), the large user base of active customers over twitter have been helping in gathering more data. Therefore, this research will expect a better sentiment analysis. Sentiment Analysis has been a challenging task that involves natural processing and machine learning. Sentence level sentiment analysis has been dealing with tagging user sentences with proper sentiment polarities. Sentence level can be distributed into positive, negative and neutral class. Aspect level has been dealing with each word in the sentiment has been directed. This level of sentiment distribution has been concerning with identifying and mining features of the product from source data. As commented by Alharbi & de Doncker (2019), tweets are utilized that are ending with positive emoticons. Most of the new works have dine with the prior polarity of words for sentiment distribution at sentence level. Models can be built with under Native Bayes and Support Vector machines. Unigram and Bigram models are utilized in feature space of conjunction in Part of speech. The unigram model have been outperforming with all other models. Polarity predictions from websites as noisy labels for training a model and utilizing 1000 manually labeled tweets have been used for testing. As per view of Sharma & Kumar (2019), the utilization of syntax features of tweets, including hashtags, the link, retweets and exclamation marks have been in conjunction with prior polarity of words and Part of speech words. Bigrams and Part of speech features are not helping in this case. Kumar & Jaiswal (2020) collected data followed with similar distant learning paradigm. They used to perform several classification task by subjective and objective. In case of subjective data, tweets ending with emoticons in the same purpose are collected (Huq, Ali & Rahman, 2017). In case of objective, crawling of twitter accounts of popular newspaper, including the New York Times have been used. Data and information utilized for training and testing has been collected with the help of search queries and it has been biased. The use of data from twitter can be analysed with the help of such models. This research will help in analyzing the best possible models for generating accurate sentiment analysis data. Polarity predictions can be done from websites that can be helpful in using syntax features of tweets based on hashtags, links and retweets.

6. Methodology

Emotional analysis of Twitter data reviews is a new area that needs more attention. The collected Twitter information has been developed from scratch to prepare for cleanliness. Also, using an element selection technique, important content is removed from the entire content. Third, the information bits are marked as physically true or negative tweets to configure the preparation package. Finally, preparation sets marked with individual highlights are provided as an aid to the classification to add additional information such as test sets.

6.1 Data Source

Determining the source of information is a critical task to advance the final investigation. Online media levels are broadly divided into three general classifications as sources of information: online

journals, miniature writing for a blog locale, and a survey website. Of all the classifications, a miniature writing for a blog website such as Twitter is ubiquitous due to its limited content and the public availability of information. The accompanying findings on the rate of Twitter development show that Twitter is being used as a site for review research.

6.1.1 Twitter Growth Rate Statistics

Approximately 6,000 tweets are posted on Twitter every second. Receive more than 350,000 tweets at any time and 500 million tweets every day. He posts about 200 billion tweets a year. Based on Twitter's experience, the number of tweets increased from 5,000 tweets per day in 2007 to 500,000,000 tweets per day in 2013, which is roughly six significant degrees. Halfway through, it posted 300,000 tweets a day in 2008, 2.5 million tweets a day in 2009, 35 million tweets in 2010, and 200 million tweets a day in 2011. Also 340 million tweets a day. For example, on March 21, 2012, six years after the launch of Twitter. Twitter no longer uses these criteria for our navigation.

6.1.2 Twitter Research

According to the current work, investigations conducted on Twitter relate to health care, advertising, legal issues, market advertising, games and so on. Phonetic or psycholangistic expertise, oral fogs and histograms. In addition, Twitter is the most promising point for investigations such as the network or outreach, content disclosure, marketing and business prospects, offer scope and the tweet ordering.

6.1.3 Tweets

The message posted on Twitter is called a tweet, limited to 140 characters. Tweets typically include the following: text, contacts, emoticons, and images. The six-second video was added in 2012 as a tweet section. With these elements in mind, extraction is applied to text, additions, pictures, emoticons or emojis, and recordings. The tweets contain three records, including the hashtag (#), retweet (RT) and record ID (@).

6.2 Data Pre-processing

Information from Twitter is difficult. Collected Information Approximately, you must prepare or swap up raw data to execute a classification. Completed works include uniform packaging, hashtags and other Twitter documents (@, RT), emoticons, URLs, stopwords, slag word decompression and extended word stress.

6.3 Feature Extraction

Prefabricated datasets have different individual properties. In the component extraction technique, we separate the different angles, descriptive words, action words, and objects, and then identify these contexts as default or negative to indicate the intensity of the whole sentence.

6.4 Sentiment Classification Technique

Generally, there are two ways to distinguish between content gradients: information-based procedures and AI procedures. Information-based procedures are also known as lexical-based strategies. The vocabulary put together a procedural center for inferring emotion-based vocabularies from the content and then recognizing hints from those vocabularies. A dictionary is a combination of well-known collected conceptual terms. This methodology is further divided into dictionary-based and corpus-based methodologies. The dictionary-based methodology searches for words placed in the ranking and then parse the word references to collect word equivalents and counterwords. The corpus-based methodology creates a summary of qualification words and then discovers additional relevant emotional words within the huge corpus, keeping in mind that they set a clear direction. To guide the approach of the dictionary, the light placement of the words that describe the grade is physically collected together with the instructions that are referenced as a means of preparatory tasks. This set is then continually developed by searching for word equivalents and counterwords in commonly used and recognized lexical word reference devices, such as WordNet or Sentiful. The basic objective of AI techniques is to build calculations that update the execution of the framework by preparing information such as models, past information, and encounters. AI gives the answer to the hypothetical order question in two consecutive previews.

1) Develop and train the model using the preparation of ensemble information, such as the actual named information.

2) Classify unlabeled or unclassified information according to the prepared or provided template.

Machine learning methods can be divided into monitored and unsupervised methods. Because we are dealing with subjective data, commonly monitored the machine learning techniques are used to analyze emotions. The monitored machine learning methods, unlike non-monitored machine learning methods, rely heavily on learning data that has already been tagged as data. Based on the training data provided, the classifier classifies the rest of the data, the test data. Sentiment analysis uses a number of monitored machine learning algorithms such as logistic regression, naive Bayesian algorithms, decision trees, support vector machines (SVMs), random forests, maximum entropy, and Bayesian networks. Choosing the right algorithm for the data and domain you choose is an important step.

7. Result and Discussion

7.1 Problem statement

Sentiment analysis done on microblogging is a new research area and contains more area to be discovered. As per view of Srivastava, Singh &Drall (2019), many work and studies are done on these research areas which are user reviews and posts over the web. Still, analysis are different from Twitter as twitter include the limitations of 140 characters per post. Therefore, users have to post their feelings and thoughts in a short manner (Emadi&Rahgozar, 2020). A proper technique can be done by implementing learning techniques which include Native Bayes and Support Vector machines. This has been expensive and research gaps are existing in such techniques. So, this research will focus on proposing a model for examining sentiments of twitter user using efficient classification techniques.

7.2 Expected Impact

The research has based on sentiment analysis related to microblogging. It is expected that this study will impact on working with unigram models in order to improve models by including data related to the closeness of the word with negation word. This study will help in specifying the window for the word that is under consideration and effect of negation involved in the model. It has been expected that closer the negation word on to the unigram model, it would affect the polarity. This study will help in identifying the POS separately from unigram models. It has been expected that identification of data based on the relative position of word in tweet affects performance of polarity of classifier.

7.3 Solution Approach

The study will focus on developing a tree that will represent tweets in order to combine many categories of features. A partial tree kernel will be used for calculating the similarity between two trees. As per view of Arora &Kansal (2019), a PT kernel has been helping in the measuring similarity between both trees by comparing sub-trees. This kernel of general class is utilized for comparing abstract objects like strings. The calculation will be done computationally by Dynamic programming techniques. Solution might be obtained by following approach that starts with initializing main tree as root. Tokenization of each tweet as in case token has been targeted as negative word and addition of leaf node in the root with the respective tag need to be done. As per view of Alsaeedi& Khan (2019), an English token needs to be mapped to its Part of speech. After that calculation of priority polarity will be done. The PT kernel tree helps in creating possible sub trees and help in comparing them together. These sub trees need to be involved in non-adjacent branches. Therefore, based on this hypothesis can be derived as follows:

H0: Bigram model can be a better model for accuracy in sentiment analysis than other models.

H1: The bigram model cannot be a better model for accuracy in sentiment analysis than other models.

Conclusion and Future work

In this article, we first go through a detailed process of ending the emotional analysis cycle to categorize Twitter's highly unorganized information into positive or negative categories. Second, we discussed several techniques to destroy emotion in Twitter data, including Twitter knowledge-based strategies and machine learning strategies. In addition, we presented the parametric correlation of the discussed monitored methods of machine learning based on our defined parameters. The various strategies for examining the

conjecture have found the domain and language to be clear. Future opportunities in opinion analysis will therefore include improvements in the technique for ordering sentiment that can be applied to all data with less focus on space. In addition, the linguistic diversity of web-based media information is a central issue that must be avoided in the future. In addition, some of the major challenges in regular language preparation (NLP) can be used to advance to the final exam, such as: Recognition of hidden or veiled emotions, recognition of parody, comparison or association management, and emoticon recognition.

References

[1]. Alharbi, A. S. M., & de Doncker, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54, 50-61.

[2]. Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2), 361-374.

[3]. Arora, M., &Kansal, V. (2019). Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. Social Network Analysis and Mining, 9(1), 12.

[4]. Emadi, M., &Rahgozar, M. (2020). Twitter sentiment analysis using fuzzy integral classifier fusion. Journal of Information Science, 46(2), 226-242.

[5]. Hassan, S. U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., ... & Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. Knowledge-Based Systems, 192, 105383.

[6]. Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. Concurrency and Computation: Practice and Experience, 32(1), e5107.

[7]. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., &Stoyanov, V. (2019). SemEval-2016 task 4: Sentiment analysis in Twitter. arXiv preprint arXiv:1912.01973.

[8]. Rodrigues, A. P., &Chiplunkar, N. N. (2019). A new big data approach for topic classification and sentiment analysis of Twitter data. Evolutionary Intelligence, 1-11.

[9]. Ruz, G. A., Henríquez, P. A., &Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92-104.

[10]. Saad, S. E., & Yang, J. (2019). Twitter sentiment analysis based on ordinal regression. IEEE Access, 7, 163677-163685.

[11]. Sailunaz, K., &Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. Journal of Computational Science, 36, 101003.

[12]. Sharma, S., & Kumar, S. (2019). Sentiment analysis on twitter posts using hadoop.

[13]. Srivastava, A., Singh, V., &Drall, G. S. (2019). Sentiment Analysis of Twitter Data: A Hybrid Approach. International Journal of Healthcare Information Systems and Informatics (IJHISI), 14(2), 1-16.

[14]. Terán, L., &Mancera, J. (2019). Dynamic profiles using sentiment analysis and twitter data for voting advice applications. Government Information Quarterly, 36(3), 520-535.