IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

Behavioural Analysis of Tweets using HFIDC Algorithm in Social Media

R. Adaikkalam[1*], A. Shaik Abdul Khadir[2]

[1,2] Dept. of Computer Science, Khadir Mohideen College(Bharathidasan University) ,Adirampattinam, Thanjavur (Dist), Tamil Nadu, India.

[*]*Corresponding Author: try2ad@gmail.com, Tel.: +91-7708198821*

Available online at: www.ijcseonline.org

Accepted: 11/Oct/2018, Published: 31/Oct/2018

***Abstract*—**Today, almost everyone is part of a socialized media, whether to express opinions on any of the products of others, business organizations, industry, educational institutions, etc., so that these views or views fluctuate they are analyzed and compared with the dictionary With the help of classifying in order to better understand whether the person who is commenting on is conducive to a positive side or a negative side and may not even support either party (neutral). Basically, the sentiment analysis is where the subjective information is extracted from the original data. The popularity of Internet users and the rapid development of emerging technologies are in parallel; active use of online commentary sites, social networks and personal blog to express their views. Through natural language processing and machine learning, with other methods for working with a lot of text along the tools, you can begin to extract the sentiments of social networks. In this article, we have discussed some of the sentiments of extraction techniques; some have taken to respond to these challenges, our approach to analyze the sentiments of social networking methods.

***Keywords*—**Behavioural analysis, Clustering, Sentiment Analysis, Social media, Tweet.

I. INTRODUCTION

Social media technologies come in many different forms, including magazines, web forums, blogs, social blogs, Weibo, wikis, social networks, podcasts, photos or images, videos, ratings and social bookmarks. Public and private opinions on a variety of topics are constantly being expressed and disseminated through many social media, where twitter is the most fashionable.

Emotional analysis is intended to determine the attitude of a speaker or writer in terms of the overall background polarity of some subjects or documents. The basic task of emotional analysis is to determine the polarity of the text in a given document, sentence, or trait / level - if it is assumed that the entity in the document, phrase, or feature / aspect is positive, negative, or neutral.

The emergence of advanced emotional categories, "transcendence", for example, in the emotional state of "anger", "sad" and "happy".

In order to measure the judgment of the media citizens, we recommend the use of emotional analysis, which has recently received the attention of business and sociologists, but still lacks a comprehensive and critical debate approach. This method is based on natural language processing, text analysis and computational linguistics methods, and measure the emotional direction towards the subject prayer. This gives the indications "based on emotional answers" through the legitimacy of its ordinary citizenship organization.

The democratization of the Internet creates content that allows a number of new technologies, media and communication tools, and ultimately leads to the emergence of social networks and the availability of text messages available in informal text. Like Twitter microblogging, blogs as LiveJournal, social networks such as Facebook and instant messaging tools such as Skype and WhatsApp are now commonly used to share ideas and comments on anything around the world, along with old mail phone text messages. The proliferation of social media content has created new opportunities to study public opinion, with Twitter being particularly popular because of its size, representation, discussion of topics and their easy access to public information.

II. RELATED WORK

Sunny Kumar et el. R input language is powerful and suitable for implementing data extraction and data analysis tools. However, when the data size is in other words, the size of the data exceeds the size of the physical memory environment R, and R gives a poor result, sometimes ending the R-Talk [1].

Gayathiri.R at el. Based on the assessment of the views of people who only consider the high accuracy of the positive

and negative evaluation rules, emotional analysis. Finally, the accuracy of the system and the accuracy of the recall measures are used to calculate the results [2].

Tian-Shyug Lee et al. The results show that in the TTE increase the number of visits, but the market optimism, the annual decline in the tourists. PEO should be able to quickly make quality decisions, accurately meet the needs of visitors, show the social media platform, constantly monitor how they are in the social media investment in who broke into ROI [3]. AmolPatwardhan et al. Detection group and the environment in the crowd spontaneous emotion. Edge detection uses a grid overlay with lines to extract features. The movement of the feature from the aspect of the movement of the reference point is used to filter through the color channel image sequence. In addition, video data collection was carried out by viewing spontaneous emotions in the participants of the sporting event. The method is independent of vision and obstruction, the results are not subject to multiple people expressing various emotions chaotically exist [4].

Anne Veenendaal et al. Check the use of color and depth data (RGB-D) motion analysis and frame recognition of human behavior. Specifically, the identification of attacks such as throwing, kicking, punching, using 3D depth sensors threatens aggression. The RGB-D data obtained from the operation and the infrared detection of the device is recorded. And it was asked to take 23 students to take positive measures. The SVM classification uses the training from the series of frames and the functional life of the combat scene based on the simulation of the combat scene. The results show that the performance of single individual stocks is better, but the system performance of the poor performance of the detection group activities [5].

Mustofa Kamal et al. describes the system to analyze the Indonesian people's views on the ASEAN Free Trade Area and extract useful information about the sensory analysis of his opinion. Our system consists of five calculation steps: (1) data gathering, (2) mining, (3) sentiment analysis, (4) normalization of time and (5) context processing. We use a new approach, based on the analysis of time and space to explore the views of public opinion on the views of the ASEAN Free Trade Area. This method automatically processes and extracts textual data to obtain information on the mood phrases. The results show that the simple method of effectiveness to obtain information on the views of people and some stakeholders and who are interested in the ASEAN Free Trade Area students have already felt that this solution has been evaluated [6].

Shaohua Wan et al. The proposed framework is expected to maximize the performance criteria for the repeated establishment of a specific gold standard for the given gold standard, and then refine the gold standard based on performance indicators. At the same time, to relieve the potential overlap of the geometrical features and the appearance of the different facial expressions, repeat the distance between our update and the performance score of the new estimated scorer [7].

J.F. Cohn et al. Head movement is mainly limited to the plane of the image. The use of high-back chairs accompanied by spontaneous smiles and smiling conditions may only have limited movement of the aircraft, and a few occur from the analysis [8].

III. PROPOSED WORK

A. Tweets Data and Performance Metrics Tweets collections

Two collections of tweets are used in the experiments to simulate targeted twitter streams. The Singapore based user tweets published in June 2010 has monitoring geo locations in targeted twitter collected from the initial collection SIN. The user group to be monitored first filled out the latest 1000 Twitter users from Singapore, with the most http://twitaholic.com fans, and then expanded the list, including fans and friends of the most important users. On Twitter in two jumps. There are a series of real-life events during the data collection period, such as the Orchard Road (Singapore Premium Shopping Belt), the 2010 FIFA World Cup and the 2010 WWDC. SIN collection contains 4,331,937 tweets.

The second collection (SGE) collects specific Twitter traffic through a monitoring group with the Singapore General Election 2011 like SIN to collect predefined keywords to simulate specific events, and only publishes tweets compiled from Singapore for user data collection. On April 13, 2011, and ended on May 13, 2011, covering the election time of 2011 Singapore (2011 and voting day April 27, nominated one day, on May 7th 2011). The SGE collection contains 226,744 tweets. It has been observed that by collecting Weibo-based users, the topics covered in the SIN collection are essentially different. In the collection of SGE coverage, on the other hand, the topics are more focused, as most of the discussion is about elections. Another inspection is that Twitter users are more proper in political discussions than improper discussions. In other words, SGE tweets are more proper than SIN.

B. Stop word selection

Stop words are most used in English, including word pronouns such as "I, him, her" or words such as "a, an, the" or prepositions. The Information Retrieval (IR) system first introduced the concept of stop words. For most of the frequency of text size, a small number of words are calculated in English. It has been pointed out that the pronouns and prepositions mentioned are not index words used to retrieve documents. Thus, it was concluded that such

words did not carry significant information about documents. Thus, the same interpretation was given stop words in text mining applications as well. In order to reduce the size of the feature values, standard practice of removing stop words from the feature values is mainly used.The stop word list that is considered to be removed from the feature space generic stop words list which is application independent. This can have a detrimental effect on the application of text mining because certain words depend on domains and applications.

C. Stemming algorithm

Stemming is the process of reducing the derivative words to derive the word, base or root form. The derivation process is called fusion. They are often referred to as stemming or stemmer algorithms. For example, an English reader recognizes the string "ACCEPTED" based on the root "ACCEPT".

D. Levenshtein algorithm

Levenshtein distance is a measure of similarity between two words. The two words are referred to as source word and target word. Between two words distances are calculated where the distance are considered about the number of insertions, substitutions or deletions required. For ex. If source word(s) is "TEST" and target word(t) is "TEST" the distance(s,t) = 0. The greater the Levenshtein distance, the more different the strings are. The Levenshtein distance is named after the Russian scientist Vladimir Levenshtein, designed the algorithm in 1965. The Levenshtein distance algorithm is also used.

- Plagiarism detection
- Spell checking
- Speech recognition

E. The algorithm description

Step 1: Set n to the length of s. Set the length of m or t. If n = 0, then my exit is returned. If m = 0, return n and exit. Construct a matrix containing 0..m rows and 0..n columns.

Step 2: Initialize the first line to 0..n. Initialize the first column to 0..m.

Step 3: Check each character of s (i from 1 to n). Check each character of t (i from 1 to m).

Step 4: If s [i] is equal to t [j], the cost is 0. If s [i] is not equal to t [j], the cost is 1.

Step 5: Set the cell d [i,j] of the matrix to the minimum value equal to: a. The next higher unit plus 1:d [i-1,j] +1. second. The cell on the left is added with 1:d [i,j-1] +1. this way. Add cost to the diagonal up and left cells: d [i-1,j-1] + cost.

Step 6: After completing the iterative steps (3, 4, 5, 6), find the distance I in cell d [n, m].

F. Performance Metric

Performance metrics used throughout the experiments include: Precision(Prec), Recall(Recall), and F1. Prec quantifies the percentage of the extracted phrases that are true named entities. Recall quantifying the percentage of correctly named entities that are correctly identified.

IV. HYBRID FREQUENT ITEMSETS WITH DOCUMENT CLUSTERING (HFIDC)

Due to its wide applicability in areas such as web mining, search engines, information retrieval and topology analysis, document grouping has been studied in depth. Unlike document classification, tagged documents are not available in document groupings. Although standard clustering techniques such as k-means can be applied to document grouping, they generally do not meet the special requirements of grouped documents: high dimensionality, high data volume, easy navigation, and meaningful clustering tags. In addition, many existing document grouping algorithms require the user to specify the number of clusters as input parameters and not sufficient to process different types of document sets in a real world environment. For example, in some document sets, clusters range in size from a few to a few thousand documents. This change greatly reduces the clustering accuracy of some of the latest generation algorithms. The intuition of our grouping standard is that each group (topic) in the document set has some common set of elements, and different groups share several sets of frequent elements. A frequent set of elements is a set of words that appear together in the smallest part of the document in the cluster. Therefore, a set of frequent elements describes the commonality of many documents in a cluster. In this technique, frequent itemsets are used to build clusters and organize clusters in the topic hierarchy. Here are the features of this approach.

- Dimensions are reduced. This method only uses frequent elements that appear in the smallest part of the document in the document vector, which greatly reduces the dimensions of the document set. Experiments have shown that packets with reduced dimensions are significantly more efficient and scalable. This decision is consistent with the linguistic study (Longman Lancaster Corpus), which covers 80% of English written text in just 3,000 words, and the results are consistent with Zipf's law.
- High clustering accuracy. The experimental results show that the proposed FIHC method overcomes the optimal document grouping algorithm in terms of accuracy. It is powerful even when applied to large and complex document sets.
- The number of clusters is an optional input parameter. Many existing grouping algorithms require the user to specify the required number of clusters as input parameters. FIHC only considers it as an optional input parameter. Even if the value is unknown, near-optimal packet quality can be achieved.

A. Performance of Tweet Segmentation

Tweet segmentation is used to extract the named entity candidates from tweets, or in other words, to identify the correct boundary of potential named entities in tweets. It is a critical component because the performance of TwiNER is heavily affected by the effectiveness of tweet segmentation. Two stickiness functions are defined by using two collocation measures, PMI and SCP, for tweet segmentation. The tweet segmentation algorithm described in Section 3 also incorporates an external knowledge base Wikipedia. Further, we normalize the segment length to favor long named entities. In this section, we study the impact of the collocation measures (PMI or SCP), the Wikipedia dictionary (Wiki), and the length normalization (Norm), based on the ground truth in SIN_g and SGE_g. We use tweet segmentation with only PMI or SCP measures as the baseline. We measure the percentage of named entities that are correctly extracted (i.e. split as a segment) as the performance metric, which is denoted as Prec as well. The experimental results are listed in Table 1.

Table 1. Performance Measures Computational Time for Tweet segmentation

No. of Tweets	Precision Time (in sec)	Recall Time (in sec)
50	1.183	1.542
100	2.569	2.896
150	3.777	4.813
200	6.13	7.15
500	9.76	10.54

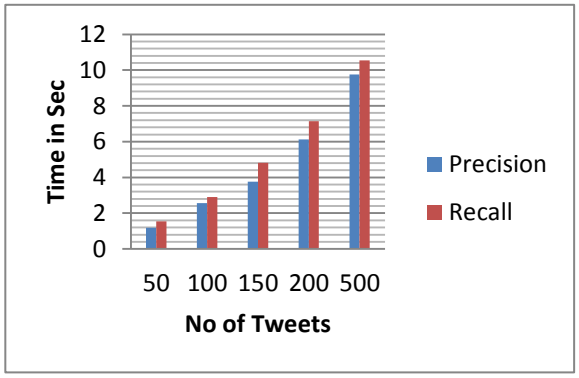

Figure 1. Performance of Tweet Segmentation

V. CONCLUSION AND FUTURE SCOPE

The results show that the combination of several polarity classifiers allows the improvement of the base classifiers. This result encourages us to continue studying the most adequate way to combine the classification power of different methodologies. Our future work will be focused on the analysis of the resolution of ties in the voting system ensembles formed by diversified components specially if these come from different information sources, such as textual data, emoticons, and lexicons can provide state-of-the-art results for this particular domain. We also compared promising tweets (i.e, word packages and feature hashes) and demonstrated their strengths and weaknesses. It turns out that the hash of features is a good choice in the emotional analysis of tweets, where computational work is the most important. However, when the focus is on accuracy, the best choice is bag-of-words. Although our results were obtained from Twitter data (one of the most popular social networking platforms), we believe that our research is also relevant to other social network analysis.

REFERENCES

[1] Sunny Kumar and Paramjeet Singh, "Sentimental Analysis of Social Media Using R Language and Hadoop: Rhadoop", 5thInternational Conference on Reliability, Infocom Technologies and Optimization (ICRITO), 2016.

[2] Gayathiri.R and Arunkumar.A, "Opinion Mining On Traffic Dataset Using Rule Based Approach", IJCSMC, Vol. 5, Issue. 3, March 2016, pg.512 – 516.

[3] Tian-Shyug Lee and Ben-Chang Shia, "Social Media Sentimental Analysis in Exhibition's Visitor Engagement Prediction", American Journal of Industrial and Business Management, 2016, 6, 392-400.

[4] AmolPatwardhan by "Edge Based Grid Super-Imposition for Crowd Emotion Recognition", Computer Vision and Pattern Recognition, 2016.

[5] Anne Veenendaal, Eddie Jones, Zhao Gang, Elliot Daly, SumaliniVartak, Rahul Patwardhan by "Fight and Aggression Recognition using Depth and Motion Data", 2016.

[6] Mustofa Kamal, Ali RidhoBarakbah, NurRosyidMubtadai by "Temporal Sentiment Analysis for Opinion Mining of ASEAN Free Trade Area on Social Media", Knowledge Creation and Intelligent Computing (KCIC), 2016.

[7] Shaohua Wan and J.K. Aggarwal, "Spontaneous facial expression recognition: A robust metric earning approach", Computer Vision Research Center, The University of Texas at Austin, Austin, TX 78712-1084.

[8] J.F. Cohn and K.L.Schmidt," The Timing of Facial Motion In Posed And Spontaneous Smiles", international journal of wavelets, multiresolution and information processing, 2, 1-12.

[9] Dileep M R and AjitDanti, 'Two Level Decision for Human age prediction using Neural Network", International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 5 Issue ICICC (May 2015).

[10] Rishi Gupta and Dr. Ajay Khunteta, "SVM Age Classify based on the facial images", International Journal of Computing, Communications and Networking, volume 1, No. 2, September-October-2012.

[11] HlaingHtakeKhaung Tin, "Perceived Gender Classification from Face Images", I.J. Modern Education and Computer Science,2012, 1, 12-18.

[12] M. Kirby and L. Sirovich, "Application of the Karhunen-Lokve Procedure for the Characterization of Human Faces", IEEE Transactions On Pattern Analysis And Machine Intelligence. VOL. 12, NO. I, JANUARY 1990.

[13] Damian Borth, RongrongJi, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors

using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia. ACM, 2013.

[14] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih -Fu Chang. Object-based visual sentiment concept analysis and application. In Proceedings of the 22nd ACM International Conference on Multimedia. ACM, 2014.

[15] YannLeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE CVPR, 2014.

[17] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li. and J. Luo, "Aesthetics and emotions in images," Signal Processing Magazine, IEEE, vol. 28, no. 5, 2011, pp. 94-115.

[18] S. Siersdarfer. E. Minack, F. Deng, and J. Hare. "Analyzing and predicting sentiment of images on the social web," in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 715-718.

[19] D. Borth. R. Ji, T. Chen, T. Breue!, and S.-F. Chang, "Large-scale visual sentiment ontology and detectars using adjective noun pairs," in Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 223-232.

[20] L. P. Marency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in Proceedings ofthe 13th international conference on multimodal interfaces, 2011, pp. 169-176.

