International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 10, October-2018

A SCALABLE METHOD OF SOCIAL MEDIA FOR EVENTS CLASSIFICATION AND MINING TWITTER DATA

V. S. S. Harshavardhan Reddy[1], Venkata Sai Chowdary Velagapudi[2]

[1]B.Tech, Dept of CSE, K L DEEMED TO BE UNIVERSITY
[2]B.Tech, Dept of CSE, K L DEEMED TO BE UNIVERSITY

Abstract: Social media is one of the mainstream sources for information retrieval and furthermore it is being utilized for sharing everyday occasions of our lives and the occurrences which are going on around the globe. Twitter can be utilized as microblogging administration used to find occasions and news progressively from anyplace in the World. Twitter posts are for the most part short and can be produced continually, so we can state that they are appropriate sources of spilling data for supposition mining and sentiment extremity location. Twitter posts on a specific occasion or subject can assist us with knowing feelings of individuals about that specific occasion or subject. Sentiment examination on twitter posts can enable us to know how individuals respond to a specific occasion and how their assessment can change if something bizarre occurred. Sentiment examination helps in numerous business regions to know surveys of an occasion. While doing any Machine Learning undertaking, the fundamental concern is an exactness of a model. In the event that the dataset is exact, at that point we can get a higher exactness of Machine Learning model. The goal of this paper is to examine approach which gives more precision of the machine learning assignment to discover sentiment extremity on Twitter data.

Keywords: data analysis, social media, Twitter, classification,

1 INTRODUCTION

The developing phenomena of social media, for example, Facebook, Twitter, Linkedin, and Instgram, with everyone has its very own attributes and its uses, are continually influencing out social orders. Facebook, for instance, is considered as a social system where everybody in the system has a responded association with another in a similar system. The relationship for this situation is undirected. Alternately, in Twitter everybody in the system does not really have a responded association with others. For this situation, the relationship is either coordinated or undirected.

In this paper, we center on twitter for data investigation, where twitter is a web based systems administration benefit that empowers clients to send and read short 140-character messages called "tweets" [1]. Notwithstanding its exposure, twitter is open for unregistered clients to peruse and screen most tweets, not at all like Facebook where clients can control the protection of their profiles. Twitter is likewise a huge social systems administration microblogging website. The monstrous information given by twitter, for example, tweet messages, client profile information, and the quantity of adherents/followings in the system assume a noteworthy job in data examination, which consequently make most investigations explore and inspect different investigation methods to get a handle on the ongoing utilized advances.

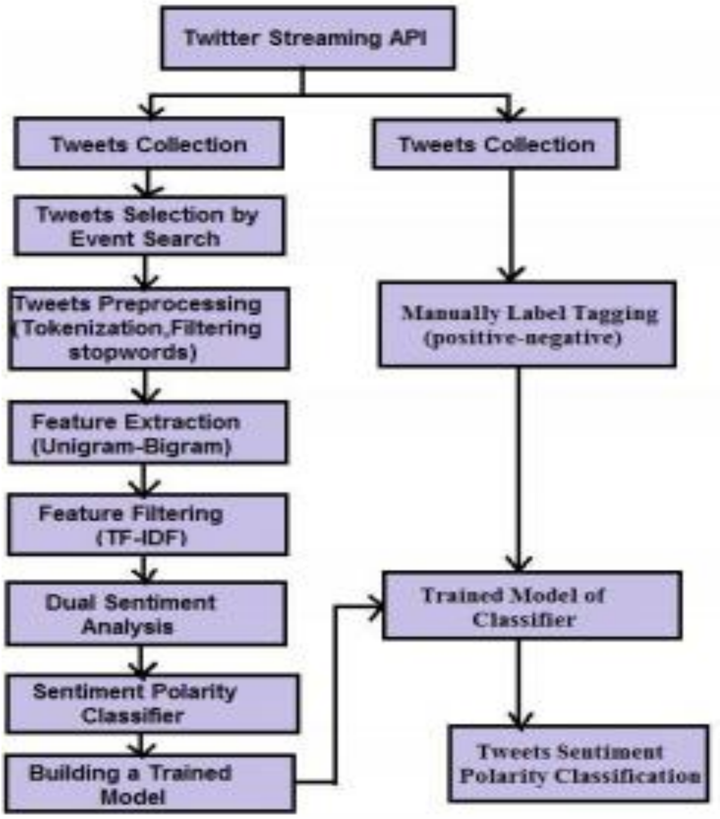
Social systems have reformed the manner by which individuals impart. Information accessible from social systems is gainful for investigation of client supposition, for instance estimating the input on an as of late discharged item, taking a gander at the reaction to arrangement change or the pleasure in a progressing occasion. Physically filtering through this data is dull and conceivably costly.

Sentiment investigation is a generally new zone, which manages extricating client feeling naturally. A case of a positive sentiment is, "normal dialect preparing is fun" on the other hand, a negative sentiment is "it's a ghastly day, I am not going outside". Target writings are regarded not to express any sentiment, for example, news features, for instance "organization racks wind area designs".

There are numerous manners by which social system data can be utilized to give a superior comprehension of client feeling such issues are at the core of Natural Language Processing (NLP) and data mining research.

In this paper we present an apparatus for sentiment examination which can investigate Twitter data. We demonstrate to naturally gather a corpus for sentiment examination and assessment mining purposes. Utilizing the corpus we manufacture a sentiment classifier that can decide positive, negative and target sentiments for an archive.

2. SENTIMENT ANALYSIS ON TWITTER DATA SYSTEM OVERVIEW


Twitter Streaming API
Tweets Collection
Tweets Collection
Tweets Selection by Event Search
Tweets Preprocessing (Tokenization,Filtering stopwords)
Manually Label Tagging (positive-negative)
Feature Extraction (Unigram-Bigram)
Feature Filtering (TF-IDF)
Dual Sentiment Analysis
Trained Model of Classifier
Sentiment Polarity Classifier
Building a Trained Model
Tweets Sentiment Polarity Classification

1. Tweets collection: This progression includes gathering the required number of tweets from Twitter API for the formation of preparing set. The data is in JSON design as an arrangement of archives.

2. Tweets pre-processing: To discover sentiment examination the significant advance is sifting through all the commotion and trivial images that don't add to a tweet sentiment from the first content. The procedures like tokenization, expelling stop-words, stemming, and so on likewise perform amid tweets pre-processing.

3. Feature extraction and feature filtering: This progression includes choosing valuable rundown of words as an element of content and evacuates an extensive number of words that don't add to the content sentiment. The idea of Ngrams is utilized where an arrangement of consecutive words is utilized for discovering the recurrence of words. After these procedures, we apply calculation of Dual Sentiment Analysis to build the precision of a model.

4. Trained model of classifier: After processing every one of the data, the last outcome is a prepared model which is connected to test data for checking the exactness of a model. A prepared model is worked by utilizing an arrangement strategy.

3. CLASSIFICATION ALGORITHMS

a) Naïve Bayes Classifier:

The Naive Bayes Classifier technique depends on the so called Bayesian hypothesis and is especially suited when the dimensionality of the data sources is high. Regardless of its straightforwardness, Naive Bayes can regularly beat more refined order techniques.

The essential component of Naive bayes classifier is finished by checking the recurrence of words identified with sentiment in the message. The tweets are grouped and scored by the quantity of matches to the sentimental words. The heaviness of hubs is balanced by the significance of tweets and more precise aftereffect of characterized sentiments can be produced.

To show the idea of Naïve Bayes Classification, consider the precedent showed in the outline above. As showed, the articles can be named either GREEN or RED. Our errand is to order new cases as they arrive, i.e., choose to which class mark they have a place, in light of the as of now leaving articles.

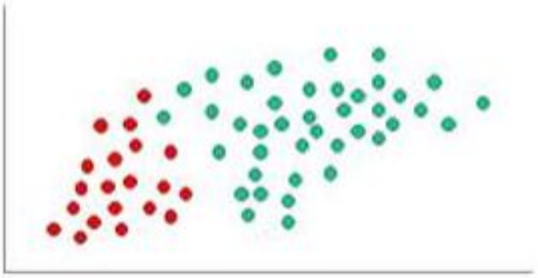


Fig. 1. Naïve Bayes Classification

Since there is twice the same number of GREEN questions as RED, it is sensible to trust that another case (which hasn't been watched yet) is twice as prone to have participation GREEN instead of RED. In the Bayesian examination, this conviction is known as the earlier likelihood. Earlier probabilities depend on past involvement, for this situation the level of GREEN and RED items, and regularly used to anticipate results before they really occur.

The Naïve Bayesian classifier fills in as pursues: Suppose that there exist an arrangement of preparing data, D, in which each tuple is spoken to by a n-dimensional component vector, X=x1,x2,..,xn, showing n estimations made on the tuple from n traits or highlights. Expect that there are m classes, C1, C2, ..., Cm. Given a tuple X, the classifier will foresee that X has a place with Ci if and just if: P (Ci | X) >P (Cj | X), where I, j ∈ [1, m] andi≠ j. P (Ci | X) is processed as:

$$P(C|X) = \prod_{k=1}^{n} P(x_k|C_i)$$

Where Xi are support vectors, Xj are testing tuples, and γ is a free parameter that uses the default value in our experiment.

b) Support Vector Machine:

SVM is for the most part utilized for content arrangement. SVM gives best outcomes than Naive bayes calculation if there should be an occurrence of content order. The fundamental thought is to discover the hyperplane which is spoken to as the vector w which isolates record vector in one class from the vectors in different class.

Support Vector machines speak to a straight model classifier. Support vector machines (SVM) are a technique by which we can order both direct and nonlinear data. For directly indivisible data, the SVM scans for the straight ideal isolating hyper plane (the straight piece), which fills in as a limit for the choice that isolates data of one class from alternate class. Numerically, an isolating hyper plane can be composed as: $W \cdot X + b = 0$, where Wa weight vector and$W$= w1, w2, ...wn. X is a preparation tuple. b is a scalar. To advance the hyper plane, the issue basically changes to the minimization of $\|W\|$, or, in other words as: $\alpha i\ n\ i$=0 $yixi$ where αi are numeric parameters, and yi are marks dependent on support vectors, . That is: ifyi=1, at that point $wi\ n\ i$=0 $xi \geq$ 1; if yi=-1 then $wi\ n\ i$=0 $xi \geq -1$[10] [11].

Furthermore, for straightly indivisible data, the SVM utilizes nonlinear mapping to change the data into a higher measurement. It in this manner takes care of the issue by finding a direct hyper plane. Functions to perform such changes are called kernel functions. The piece function chose for our trial is the Gaussian Radial Basis Function-(RBF):

$$K(X_i, X_j) = e^{-\gamma\|X_i - X_j\|^2 /_2}$$

4. LITERATURE REVIEW

With the end goal to present what is prescient investigation we have to know where it originates from. Hence, initially the umbrella that covers those frameworks is called business intelligence (BI). Business intelligence is a mix of devices planning to improve the basic leadership in an association by changing data into helpful information and learning which is extricated by using data mining instruments and systematic techniques. BI frameworks help in investigating and enhancing association's execution, making new procedures to upgrade the income and benefit of the association. In this manner Data mining is a piece of Business intelligence functionalities as characterized by Gartner who portrayed BI as a product stage conveying 14 abilities isolated into three gatherings of functionalities including coordination, information conveyance and examination functionality which contain the data mining and prescient displaying. While data mining is considered as the computerized procedure to identify the obscure examples in the organized data of the association

[2][3]. Another [4] depicts data mining as the procedure to gather, channel, plan, examine and store data that will be utilized to make helpful information and supporting the data examination and prescient displaying.

Prescient investigation as a rule is utilized to distinguish the connections and examples in data with the end goal to anticipate the future by breaking down the past and taking better preventive choices. In this manner, the prescient investigation point of utilization contrast starting with one industry then onto the next, for example an advertiser can utilize the prescient examination to anticipate the clients' reaction to a publicizing effort, or an item vender can utilize it to foresee the development of item costs, or it very well may be utilized to distinguish patterns, for example, in banks if a supervisor wish to perceive the most beneficial clients, or alarm a Visa client to a likely fake charge. In this way the prescient examination help in noting numerous inquiries, for example, what will occur if the interest of items diminishes? Or on the other hand if providers' costs increment? What is the hazard to lose cash another business? [8]

Actually, in the training area as one of the essential parts the prescient examination have been utilized for various purposes and by utilizing diverse models and instruments. For instance, an investigation center in conduct examination in college with the end goal to foresee whether understudies are powerless against freak philosophies which can prompt psychological warfare [9]. While, another investigation demonstrates the need and advantage from utilizing the prescient examination in the instructive area which will assist the instructive foundations with increasing the maintenance of understudy and upgrading their outcomes and accomplishments. The specialists concentrated on the utilization of order calculations particularly the choice tree to show signs of improvement forecast results. The need in this investigation is that it's not particular and not tried by genuine data [10]. In a similar industry, another exploration pointed by the utilization of an execution explanatory technique to foresee the last understudies' execution in a particular course amid the semester and check the ones that will fall flat and have low execution in exams. The specialists utilized the choice tree calculation to foresee the understudies' last execution results. All things considered, this examination will be more gainful by doing further investigation of finding the understudies' errors in the exams, their learning connections with the instructive framework by utilizing diverse calculations, for example, affiliation administer calculation [11]. Also, [12] dissected the understudies' data with the end goal to foresee the drop out component of understudies and finding the principle factors that impact the open sources dropping by understudies. Along these lines, to do this investigation the analysts connected component determination calculations by utilizing WEKA instrument and afterward arrangement calculations. In this way, the outcomes will be more precise with the utilization of various calculations, for example, affiliation and bunching techniques.

M. S. Neethu and R. Rajasree, [1] Sentiment investigation manages recognizing and ordering assessments and gathering emotions or sentiment appeared in source content. Web based systems administration is making a boundless proportion of thought rich information as tweets, declarations, blog passages et cetera. End examination of this customer created information is incredibly significant in knowing the evaluation of the gathering. Twitter thought examination is troublesome appeared differently in relation to general sentiment examination as a result of the region of slang words and off base spellings. The most extraordinary limit of characters that are allowed in Twitter is 140. Information base methodology and Machine learning approach are the two techniques used for exploring appraisals from the substance. In this paper, we endeavor to research the twitter posts about electronic things, for example, mobiles, versatile PCs et cetera using Machine Learning approach. By doing sentiment Analysis in a specific space, it is possible to recognize the effect of territory data in feeling course of action. We demonstrate another segment vector for gathering the tweets as useful, opposite and focus society's sentiment about things.

G. Gautam and D. Yadav,[2] The expansive of World Wide Web has brought another technique for communicating the conclusions of individuals. It is a medium with a colossal data where customers can see the perspectives of various customers that are characterized into different sentiment classes and are developing as a key factor in decision making. This paper adds to the sentiment examination for customers' audit characterization which is helpful to break down the data as tweets where ends are unstructured and are either positive or negative. For this we first preprepared the dataset, after that isolated the modifier from the dataset that make them include vector, at that point pick the part vector list and connected machine learning based characterization estimations i.e., : Naive Bayes, Maximum entropy and SVM nearby the Semantic Orientation based Word Net which removes proportional words and comparability for the substance highlight.

J. Akaichi, Z. Dhouioui and M. J. Lopez-Huertas Perez, [3]In late years, content mining and sentiment investigation have gotten extraordinary thought due to the wealth of supposition information that exist in social systems administration locales for instance, Facebook, Twitter, et cetera. Sentiments are anticipated on these media utilizing messages for conveying feelings, for instance, kinship, outrage, cheerful, fulfillment, et cetera. Existing sentiment examination tends to recognize client conduct and perspective yet at the same time unsatisfied due to complexities in passed on writings. In this paper, we focus on the utilization of data digging for sentiment grouping .Our point is to remove helpful data, about customers supposition and conduct amid this critical period. Thus, we propose a procedure i.e., Support Vector Machine (SVM) and Naïve Bayes. We additionally build sentiment vocabulary dependent on the emojis. Moreover, we play out some comparable investigation between two machine learning estimations SVM and Naïve Bayes through a preparation demonstrate for sentiment order.

M. S. Schlichtkrull, [4] Emoticons have in the writing been appeared to alter rather than give excess to the going with literary message. Notwithstanding this, emojis are routinely utilized as mark for sentiment arrangement assignment. This paper intends to investigate the wonder and discover more striking emoji feeling relationship through an inserted based machine learning process. Utilizing important part investigation and k-implies bunching, it is indicated how comparative emojis shape bunches in vector space.

A. Celikyilmaz, D. Hakkani-Tür and J. Feng, [5]We utilize machine learning way to deal with manage sentiment grouping on twitter messages (tweets). We separate each tweet into two classes: polar and non-polar. Tweets with positive or negative sentiments are seen as polar. Generally considered as nonpolar. Sentiment Analysis of tweets can advantage for instance purchasers and showcasing analysts, for obtaining information on various item and administrations. We present techniques for content standardization of the loud tweets and their order concerning the extremity. We attempt distinctive strategies with a blend show approach for sentiment words, which are later utilized as pointer highlight of the order display. Considering a most elevated quality outfit of tweets, because of this new methodology, we get F-scores that are 10% superior to an order standard that utilizations crude word n-gram highlights.

5. CONCLUSION AND FUTURE WORK

The task "Sentiment analysis on Twitter data" intends to recognize whether a tweet is certain or contrary, it can assist us with knowing surveys of individuals about any subject/theme. It can likewise assist us with knowing whether the circumstance in some hazardous situation is basic or not. So by utilizing those outcomes, it would be simple for social foundations and government to take numerous choices for individuals. Likewise, client audits assist associations with improving in some specific territories and furthermore they would have come to know the valuation for clients for some item/thought. Additionally, our primary worry in Machine Learning Algorithm is the exactness of the outcome, so our proposed framework presents Dual Sentiment Analysis (DSA) which gives us extensive dataset. Thus, the precision of our model likewise builds which is our principle concern. As we have seen that this model can be utilized to distinguish whether the tweet is sure or negative, in the event that we bring the tweets in regards to fiasco put and in the event that we discover that the greater part of the tweets are demonstrating "negative" that implies the circumstance around there is basic, so as the future extension we can outline an algorithm which can get all the "Negative" named tweets and will discover the real issues here and will specifically advise foundations about those significant issues.

REFERENCES

[1] H. Isah, P. Trundle and D. Neagu, "Social networking media identifying for product safety using data mining and sentiment analysis," Computational Intelligence (UKCI), 2014 14th UK Workshop on, Bradford, 2014, pp. 1-7.

[2] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter using supervised learning process," (COMSNETS) Communication Systems and Networks, 2014 Sixth International Conference on, Bangalore, 2014, pp. 1-8.

[3] Tiara, M. K. Sabariah and V. Effendy, "Sentiment analysis on Twitter using lexicon-based and support vector machine methodology for assessing the performance of a television program," (ICoICT) Information and Communication Technology, 3rd International Conference on 2015, Nusa Dua, 2015, pp. 386-390.

[4] R.Rajshree and M S. Neethu, "machine learning techniques used in sentiment analysis in twitter," (ICCCNT) Computing, Communications and Networking Technologies, Fourth International Conference 2013 on, Tiruchengode, 2013, pp. 1-5.

[5] G. Gautam and D. Yadav, "Sentiment analysis in twitter using semantic analysis and machine learning approaches," Contemporary Computing (IC3), 2014 Seventh International Conference on, Noida, 2014, pp. 437-442.

[6] J. Akaichi, Z. Dhouioui and M. J. Lopez-Huertas Perez, "For sentiment classification text mining face book updates on," (ICSTCC) System Theory, Control and Computing, 2013 17th International Conference, Sinaia, 2013, pp. 640-645.

[7] Bongwon, S., Lichan, H., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting Retweet in Twitter network. Proceedings of the 2010 IEEE Second International Conference on Social Computing (pp. 177-184).

[8] Ye, S., & Wu, F. (2013). Measuring message propagation and social influence on Twitter.com. International Journal of Communication Networks and Distributed System, 11(1), 59-76.

[9] Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. International Journal of Social Research Methodology, 16(2), 91-108.

[10] Becker, H., Chen, F., Iter, D., Naaman, M., Gravano, L.: Automatic Identification and Presentation of Twitter Content for Planned Events. In ICWSM, 2011.

[11] Becker, H., Iter, D., Naaman, M., Gravano, L.: Identifying content for planned events across social media sites. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 533-542). ACM, 2012.

[12] Bekkerman, R., McCallum, A.: Disambiguating web appearances of people in a social network. In Proceedings of the 14th international conference on World Wide Web (pp. 463-470). ACM, 2005.