International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Comparative Analysis of Algorithms for Twitter Sentiment Analysis

Majid Bashir Ahmad[1], Saba Hanif[2], Kalim Sattar[3], Waseem Akram[4]
[1]Department of Computer Science, University of Lahore (Pakpattan Campus), Punjab, Pakistan
[2,3,4]Department of Computer science, COMSATS Institute of Information Technology, Vehari Punjab, Pakistan

ABSTRACT

The progress from web 1.0 to web 2.0 has empowered direct connection amongst users and its different assets and administrations, for example, social media networks. In this research paper, we have dissected algorithms for sentiment analysis which can be utilized to use this enormous data. The objectives of this paper are to gadget a method for acquiring social network opinions and separating highlights from unstructured content and dole out for each component its related estimation in an unmistakable and proficient way. In this project, we have connected naive Bayes, support vector machines and most extreme entropy for investigation and delivered an explanatory report of the three subjectively and quantitatively. We played out the task observationally and broke down the subsequent information utilizing an exceed expectations device to get comparative analysis of the three algorithms for characterization.

Keywords : Web 2.0, Social Network, LIBSVM, Support Vector Machines

I. INTRODUCTION

Coordinate collaboration in the web and the environment has prompted the accessibility of immense data in the web. A web-based social networking system, for example, tweeter, Face book, linkeldn and whatsapp has empowered individuals to impart insights real-time. Thus, they make utilization of individuals' sentiments to settle on choices for people as well as for government and business sectors. Having such mass volume of information from various data sources make it hard to take helpful and acceptable choice because of three factors. Individuals can't read the mass measure of information accessible, information on the web is unstructured; semi structured and heterogeneous in nature and data about a similar item is frequently spread over an expansive number of sites and user accounts. Besides, differential feature format and a few items utilizing distinctive names make the resulting yield of assessment mining and sentiment analysis concerning that area of the online items. The levels of ordering sentiments incorporate document level, sentence level/express level and angle/feature level. We utilize it as per the level intrigue. In our research venture, we have utilized component level since we are gathering opinions around a few parts of a similar item and inside a similar archive. We will subject the information to the three calculations naive Bayes, support vector machines and maximum entropy.

1.1 Twitter

This is an ongoing information network that associates people to the most recent stories, thoughts, sentiments and news about what you find fascinating. To take after discussions and most convincing data, you will essentially look through their accounts. Blasts of data called tweets will be found in the tweeter accounts. A tweet has 140 characters in

length however it gives a considerable measure of data to be found. You will discover photographs, recordings, and discussions specifically in the tweets to get the entire story without a moment's delay. In this venture we utilized crude tweeter information gathered from a few records utilizing the tweeter API and preprocessed with the end goal of testing.

1.2 Sentiment Analysis

Machine learning use algorithm to explain the sentiment analysis as a standard content arrangement problem that makes utilization of syntactic or etymological highlights. The grouping model is identified with the highlights in the hidden record to one of the names. The model is utilized to foresee a class name for each occurrence of obscure class. It is difficult to characterize when just a single is doled out to an occurrence.

II. METHODS AND MATERIAL

2. Supervised Learning

The supervised learning strategies rely upon the presence of marked training reports. There are numerous sorts of directed classifiers in writing. The short points of interest the absolute most as often as possible utilized classifiers in sentiment analysis.

2.1 Probabilistic Classifiers

Probabilistic classifiers utilize mixture models for grouping. The mixture shows accept that each class is a part of the mixture. Every mixture segment is a generative model that gives the likelihood of inspecting a specific term for that segment. These sorts of classifiers are likewise called generative classifiers.

2.2 Support Vector Machines

It is a straight classifier which is successful and can accomplish great execution. In high dimensional list of capabilities space. In our undertaking, it demonstrated that the classifier demonstrated the most dependable regarding accuracy, precision and precision of the supposition procedure. We prepared with LIBSVM (Chang and Lin, 2011) a broadly utilized apparatus in many considers.

2.3 Maximum Entropy

The thought behind maxent classifiers is that we ought to incline toward the most uniform models that fulfill any given limitation. Maxent models are highlight based models. Maxent makes no autonomy suppositions for its highlights, not at all like Naïve Bayes. This implies we can include highlights like bigrams also, phrases to maxent without agonizing over component covering. The guideline of most extreme entropy is helpful unequivocally just when connected to testable data. A snippet of data is testable in the event that it can be resolved whether a given dissemination is predictable with it. The significant preferences of utilizing maxent or its varieties are:

- Precision.
- Consistency – This calculation indicates consistency in comes about and if priors are utilized outcomes additionally enhance over some undefined time frame.
- Execution/Efficiency - Can deal with enormous measures of information.
- Adaptability - The calculation is adaptable of having a wide range of wrote of information in a bound together stage and group it as needs be.

III. Naive Bayes

Naive Bayes is utilized as a classifier in different true issues like Sentiment analysis, email Spam Detection, Email Auto Grouping, email arranging by need, Document Categorization and Sexually unequivocal substance identification. Innocent Bayes grouping model figures the back likelihood of a class, considering the conveyance of the words in the archive. The model works with the bows highlight extraction which disregards the position of the word in the document. It utilizes Bayes Theorem to

anticipate the likelihood that a given list of capabilities has a place with a specific mark. The real preferred standpoint of Naïve Bayes is it requires low preparing memory and less time for execution. It's exhorted that this classifier ought to be utilized when Training time is a pivotal factor in the framework. Credulous Bayes is the standard calculation for inquiries about in choice level order issue. In nearness of constrained assets regarding CPU and Memory Credulous Bayes is suggested classifier.

IV. Methodology

We led our exploration experimentally and the information comes about were led quantitatively and subjectively. The analysis of the subsequent information was finished utilizing an exceed expectations application since the informational collection was not huge. The entire procedure can be abridged as follows:

4.1 Design Architecture

- Removed tweets from web-based social networking utilizing an extraction content. Twitter API was utilized to gather tweets and after that put away in a MSQL database.
- Preprocessing and cleaning of the information.
- The information is then partitioned into 75% for preparing and 25% for test informational collection.
- Preparing the information in order to think of a model that can be utilized to group new and unadulterated tweets.
- Utilizing the model produced to group posts which include from the tweets gathered and characterizes them into the three polarities i.e. negative, positive and impartial.
- Results analysis is accomplished from the classifiers created and the conclusions drawn.

4.2 Preprocessing

Preprocessing the information is finished by cleaning and setting up the content for grouping. Online writings contain normally loads of commotion and uninformative parts, for example, HTML labels, contents and commercials. Furthermore, on words level, many words in the content don't affect the general introduction of it. Keeping those words makes the dimensionality of the issue high and henceforth the order more troublesome since each word in the content is dealt with as one measurement. To diminish the commotion in the content should help enhance the execution of the classifier and accelerate the grouping procedure, in this manner supporting continuously opinion analysis. The entire process includes a few stages: online content

cleaning, void area expulsion, extending abbreviation, stemming, stop words evacuation,

invalidation dealing with lastly include determination. Highlights with regards to feeling mining are the words, terms or expressions that firmly express the conclusion as positive or negative. This implies they highly affect the introduction of the content than different words in a similar content.

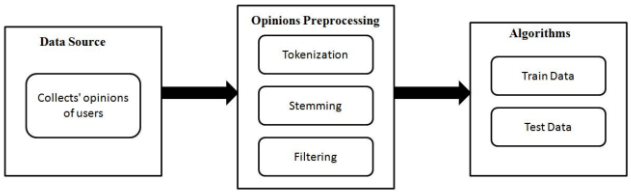

Figure 1. Implementation Model

4.3 Filtering

Re hashed words like great to demonstrate their force of articulation are disposed of as they are absent in the sentiwordnet subsequently additional letters in the word must be dispensed with. This disposal takes after the decide that a letter can't rehash more than three times.

4.4 Questions

Questions such like which, how, what and so on are not going to add to extremity consequently with a specific end goal to lessen the many-sided quality, such words are evacuated.

4.5 Removing Special Characters

Extraordinary characters like () {} [] and so forth ought to be evacuated with a specific end goal to dispose of errors amid task of extremity. For instance, "it's great" means if the characters are not evacuated may connect with the words and make those words inaccessible in the lexicon.

4.6 Removing Retweets

Many individuals may duplicate someone else's tweets and retweet utilizing an alternate account. This happens in the event that he loves another user's tweet.

4.7 Removing URLs

For the most part URLs does not add to analysis of the sentiment in casual content e.g. "I have signed into www.ecstacy.com as I am exhausted". This is negative however might be unbiased in view of the nearness of the word bliss.

4.8 Feature Selection in Sentiment Classification

Sentiment Analysis undertaking is viewed as a sentiment grouping issue. The initial phase in the SC issue is to separate and select content highlights. A portion of the present highlights are: Terms nearness and recurrence: These highlights are singular words or word n-grams and their recurrence checks. It either gives the words double weighting (zero if the word shows up or one assuming generally) or utilizations term recurrence weights to demonstrate the relative significance of highlights. Parts of discourse (POS): discovering descriptors, as they are critical pointers of opinions. Opinion words and expressions: these are words ordinarily used to express conclusions including great or awful, as or hate. On the other hand, a few expressions express sentiments without utilizing assessment words. For instance: cost me an arm and a leg. Negations: the presence of negative words may change the conclusion introduction like not great is equal to awful.

Comparative Analysis of the Algorithms

Feature	Naïve Bayes	SVM	Maximum Entropy
Accuracy	Good 71.3%	High 78%	Good 72%
Performance	Good	High	Good
Time Required for Training	Less	High	Balanced
Required Memory	Low	High	High
Easiness	Easy	Tough	Tough
Accuracy Consistency	Fluctuating	Consistent	Fluctuating

Figure 2. Comparative Analysis of the Algorithms

V. Comparative Analysis of the Algorithms

From our investigation, it was obvious that each sort of grouping model had its own difficulties. The determination of characterization models can be settled on the premise of assets, precision necessity and preparing time available. Considering the help vector machines which demonstrated that it was difficult to execute, high memory requirements, consistent in information yield and devours additional time in preparing, what's more, the classifier was best fit for use in sentiment analysis. Be that as it may it requires high preparing time and handling power this subsequently enhanced the precision of the classifier. If preparing power is an issue and memory is an issue, then the naive bayes classifier is chosen because of its low handling force and memory utilization less preparing is required time is required. If you capable handling framework and memory, then most extreme entropy ends up being a commendable option. Support vector machines ended up being normal in all angles and

accordingly ended up being the best decision for sentiment analysis.

VI. Conclusions and Future Work

In our research we applied machine learning techniques on large data set. We discuss that how to gather twitter data set for cleaning and sentiment analysis process. This process gives us algorithms accuracy results e.g SVM has 78%, Naïve Byes has 71.3% and Maximum entropy has 72% accuracy.in all of these algorithms SVN is more consistent for giving accurate results. However these results can be improved in future.

VII. REFERENCES

[1]. Kiplagat Wilfred Kiprono, Elisha Odira Abade. "Comparative Twitter Sentiment Analysis Based on Linear and Probabilistic Models". Nairobi, Kenya : School of computing and informatics, University of Nairobi,, 2016.

[2]. Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

[3]. Barbosa, Luciano and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. InProceedings of the International Conference on Computational Linguistics (COLING-2010). 2010. 17.

[4]. Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment in twitter events. Journal of the American Society for Information Science and Technology, 62(2):406-418, 2011.

[5]. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008; 2: 1-135.

[6]. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. User-level sentiment analysis incorporating social networks. Arxiv preprint arXiv:1109.6018, 2011.

[7]. Soper, D. S., & Turel, O. 2012. An n-gram analysis of Communications 2000-2010. Communications of the ACM, 55(5): 81-87. doi:10.1145/2160718.2160737.

[8]. Jennifer, D. "Affective Text based Emotion Mining in Social Media."International Journal 2.3 (2014).

[9]. Arya, Arti, et al. "A text analysis based seamless framework for predicting human personality traits from social networking sites." International Journal of InformationTechnology and Computer Science (IJITCS) 4.10 (2012): 29.

[10]. M. Hu and B. Liu, "Mining and summarizingcustomerreviews," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, (New York, NY, USA), pp.168-177, ACM, 2004.