A COMPREHENSIVE ANALYSIS OF THE SENTIMENT ON THE HASHTAG OF THE SOCIAL NETWORKING SITES- TWITTER

Swayam Jain

Modern School, Barakhamba Road, New Delhi

ABSTRACT

Opinion Mining has further developed internet shopping stages, analytical studies from political surveys, business insight, etc. By doing assessment mining in a particular region, recognizing the impact of region data in sentiment is a conceivable examination. In this, we are attempting to break down the Twitter posts about Hashtags like #MakeinIndia utilizing the Machine Learning approach. We set forth a component vector for characterizing the tweets as certain, negative and nonpartisan. I applied machine learning calculations from that point forward, particularly MaxEnt and SVM. We used Unigram, Bigram and Trigram Feature to create many features to prepare direct MaxEnt and SVM classifiers. Eventually, we have estimated the classifier's presentation as far as general precision.

INTRODUCTION

The web looks great on the off chance that it permits people to offer their viewpoint. It works in weblog posts, online discourse discussions, sites, etc. People depend on this client created content. At the point when an individual longing to purchase an item, he understands surveys on the sites. How much happiness is exorbitantly high for an individual to examine. Along these lines, robotizing is required. Opinion Analysis on Twitter is extremely hard because of its brief period. The presence of shoptalk expressions, emojis and incorrect spellings in tweets is important for a pre-processing venture before including extraction. May carry out highlight choice methodologies for social event significant abilities from the text in tweets. The component choice is acted in levels to separate pertinent highlights. In the first period of twitter, explicit capacities are removed. Then, at that point, these are eliminated from tweets to make ordinary text. Then brand extraction is utilized to get extra highlights. That is the idea utilized in this paper to produce an effective element vector for considering twitter opinions. We have made insights set by gathering tweets for a definite time frame outline [1].

There are countless sorts of AI methods utilized for assessment mining. Managed learning depends on named informational collection.

These classified realities are appointed to the model all through. These classified, named sets are talented in supplying good results. Individuals getting to know does now not envelop a class or division or type, and they never again give the right objectives in any regard, so clustering is made due. This examination paper depends on managed AI [2]. The vocabulary-based technique has a

place with unaided realizing, which no longer needs preparing informational collection and handiest rely on the word reference utilized [3]. Make named tweets that can utilize to prepare information in the Support Vector machine strategy and Maximum Entropy technique so there could be no manual name.

Twitter messages posted are easy going. Due to the anomalistic idea of casual printed content, handling or examination of such text is more troublesome. With information preprocessing, the formal and casual text is separated. Formal printed content requirements less preprocessing. The simple text incorporates emojis, unfortunate language structure, utilization of shoptalk and mockery or no word reference favoured words. So In many cases, investigation of this text classification is extreme [4].

OPINION ANALYSIS

Characterization of Text Mining is Sentiment examination, which refers to recovering related realities and nontrivial designs from the unstructured content idea. Opinion class is a straightforward task when contrasted with text auto-order. Opinion order is the parallel extremity type those arrangements with a tiny number of preparing [5].

Table I: Details of Dataset

No. of Tweets	Positive	Negative	Neutral
7000	2313	2359	2328
10000	3339	3271	3390
12000	4018	3954	4028
15000	5063	5035	4902

A. Collection of Information

We use Twitter information in our tests for improvement and preparation. We utilize the Hash Tag informational collection from Twitter API. We gathered 7000 tweets, etc., of Hash Tag [15]. We have removed tweets in English. Table I shows the quantity of Twitter messages and the dispersion across classes.

B. Pre-processing

Tokenization is utilized to partition printed content into words, images, and other significant elements alluded to as "Tokens". May isolate tokens by utilizing the utilization of whitespace characters. The standardization cycle is distinguishing expressions of shortened forms accessible in the tweet find out after which contractions are supplanted by Full importance supplant, e.g., "OMG" by way of "Wow" [16]. Assuming that the word is rehashed, the structure will eliminate these words into real significance. Erase the HTTP joins additionally and Shoptalk words like @, RT and so on and Stop words. Making parting the word into tokens that tokens in Unigram

structure. It is making Unigram word, then, at that point, eliminating stop words into the Unigram word list. This word is to assess Chi-squared. These Chi-squared passed through the classifier. These are getting Unigram Word of List. Same as Bigram and Trigram to assess Chi-squared outcome. These are Getting Bigram and Trigram's expression of rundown. The MPQA subjectivity dictionary is a word list marked with opinion extremity. We conveyed dictionary highlights content of wordlist from the vocabulary that can be addressed by certain nonpartisan and negative

SENTIMENT CLASSIFICATION TECHNIQUES

A. SVM Classifier:

SVM Classifier involves a huge wiggle room for order. It keeps a wide gap between the two classes [1]. It isolates the tweets utilizing a hyperplane. SVM utilizes the qualifications of work characterized as ' F' is the component vector,' we are the load's vector, and ' b' is the predisposition vector. φ() is the nonlinear planning from contribution to high layered highlight space.' w' and' b' are consequently educated on the preparation set. Here we involved a direct piece for order.

B. Greatest Entropy

Greatest entropy expands the entropy portrayed on the restrictive likelihood dissemination. It even handles cross-over highlights and is equivalent to calculated relapse, which tracks down appropriation over classes. It additionally follows positive trademark special case limitations [2]

Where c is the class, and d is the tweet message. The weight vectors decide the meaning of an element in characterization. It follows the comparative cycles as guileless Bayes referenced above and presents the extremity of the feelings [2].

ASSESSMENT

After doing pre-processing, all tweets are separated into two sets preparing and testing. The preparation set contains 70% of the information, and testing contains 30% of the information. Then advances followed include determination which are unigram, bigram and trigram. We find unigram, bigram, and trigram by utilizing separate preparation and testing informational collections.

Table II: Accuracy of SVM for Different size of Data sets

No. of Tweets	Unigram	Bigram	Trigram
7000	63.99	89.99	95.82
10000	64.26	91.19	96.65
12000	64.28	91.49	96.78
15000	63.31	91.02	96.77

Table III: Accuracy of MaxEnt for Different size of Data sets

No. of Tweets	Unigram	Bigram	Trigram
7000	87.17	93.75	97.53
10000	93.24	96.48	99.05
12000	94.45	96.59	99.16
15000	95.14	97.32	99.23

CONCLUSION

This paper has exhibited a strategy for naturally gathering tweets utilizing Twitter API. We pre-process and include the choice to clean the tweets utilizing the gathered tweets, i.e., eliminate redundant information from tweets. We utilize a few tweets to prepare the opinion classifier from those tweets. Classifier tracks down the positive, negative and nonpartisan opinions from tweets. The classifier is given Machine learning calculations like the SVM classifier and MaxEnt classifier that utilizes Unigram, Bigram and Trigram as their characterization highlight. By utilizing Unigram, Bigram and Trigram highlight determination alongside two classifiers, the framework got such immersion of expanding preparing information to assess SVM exactness of the mean of Unigram 63.31%, Bigram 91.02% and Trigram 96.77% results. Assess MaxEnt exactness of the mean of Unigram 95.14%, Bigram 97.32% and Trigram 97.23%.

REFERENCES

[1] Neethu M S, Rajasree R, “Sentiment Analysis in Twitter using Machine Learning Techniques”. Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on IEEE pp. 1-5, july 2013.

[2] Geetika Gautam, Divakar yadav, “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis”. 7th International Conference on Contemporary Computing on IEEE pp. 437-442, 2014.

[3] Tiara, Mira Kania Sabariah, Veronikha Effendy, “Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program”. 3rd International Conference on Information and Communication Technology (ICoICT) IEEE pp. 386-390, 2015

[4] Seyed-Ali Bahrainian, Andreas Dengel, “Sentiment Analysis using Sentiment Features”. IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT) pp. 26-29, 2013

[5] Kishori K. Pawar, R. R. Deshmukh, “Twitter Sentiment Analysis: A Review”, International Journal of Scientific & Engineering Research, Volume 6, Issue 4, 9 ISSN 2229-5518, pp.957-964, April-2015.

[6] Parisa Lak, Ozgur Turetken, " Star Ratings Versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions", 47th Hawaii International Conference on System Science, pp.796- 205, 2014.

[7] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision". CS224N Project Report, Stanford, 2009.

[8] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of LREC, Page no. 1320-1326 ,2010

[9] Luciano Barbosa and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data". In Proceedings of the 23rd International Conference on Computational Linguistics pages no 36–44, 2010

[10] Adam Bermingham and Alan F Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?". In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM, 2010.

[11] Albert Bifet and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data". In Discovery Science, pages 1–15. Springer, 2010.

[12] Dmitry Davidov, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys". In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. ACL, 2010.

[13] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. "Improved part-of-speech tagging for online conversational text with word clusters". In Proceedings of NAACL 2013, 2013.

[14] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. "The Edinburgh twitter corpus". In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 25–26, 2010

[15] Sachin Madhukar Ramteke, Sachin N. Deshmukh, "Twitter Sentiment Analysis using Adaboost Classification". International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, pages no 6444-6450, April 2016

[16] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!". Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media page no. 538-541, 2011

[17] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal (2014) 5, pp.1093–1113. 2014.