iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39201

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Sentiment Analysis on Twitter Hashtag Datasets

Ganesh K. Shinde[1], Vaibhav N. Lokhande[2], Rasika T. Kalyane[3], Vikas B. Gore[4], Umesh M. Raut[5]
[1, 2, 3, 4, 5] *Department of CSIT, Dr B.A.M. University Aurangabad Maharashtra India*

Abstract: Sentiment Analysis has improvement in online shopping platforms, scientific surveys from political polls, business intelligence, etc. In this we trying to analyse the twitter posts about Hashtag like #MakeinIndia using Machine Learning approach. By doing opinion mining in a specific area, it is possible to identify the effect of area information in sentiment analysis. We put forth a feature vector for classifying the tweets as positive, negative and neutral. After that applied machine learning algorithms namely: MaxEnt and SVM. We utilised Unigram, Bigram and Trigram Features to generate a set of features to train a linear MaxEnt and SVM classifiers. In the end we have measured the performance of classifier in terms of overall accuracy.

Keywords: Sentiment analysis, support vector machine, maximum entropy, N-gram, Machine Learning.

I. INTRODUCTION

The internet looks good if it allow human beings to express their opinion. It is work in the form of weblog posts, on-line dialogue forums, web sites etc. Human beings rely on this user generated content. When a person desires to buy a product he reads reviews on the websites. The amount of content is excessively high for a person to investigate. Therefore it is need to automate this.

Sentiment Analysis in twitter is very hard due to its short period. Presence of slang phrases, emoticons and misspellings in tweets it is necessary to do a preprocessing step before feature extraction. Feature selection strategies for gathering relevant capabilities from text which may be implemented to tweets additionally. The feature selection performed in levels to extract relevant features. In first phase twitter specific functions are extracted. Then these are removed from tweets to create regular text. Then characteristic extraction is used to get additional features. That is the concept used in this paper to generate an efficient feature vector for studying twitter sentiments. We have created a statistics set by way of accumulating tweets for a sure time frame [1].

There are so many types of machine learning techniques used for opinion mining. Supervised learning is based on labeled data set. These categorized facts are assigned to the model throughout. These categorized labelled set are skilled to supply affordable outputs in the course of decision making. Unsupervised getting to know does now not encompass of a class or division or type and that they do no longer provide with the correct goals in any respect so clustering is managed. This research paper is based on the supervised machine learning [2]. The lexicon-based method is belong to unsupervised learning, which does no longer want training data set and handiest depend upon the dictionary that is used [3]. Create labelled tweets which can be used as training data in Support Vector machine technique and Maximum Entropy method so there could be no manual labelling manner.

Twitter messages posted are casual. Due to the anomalistic nature of informal textual content, processing or analysis of such text is more difficult. With the help of data preprocessing formal and informal text is differentiated. Formal textual content needs less preprocessing. Informal text includes emoticons, poor grammar, use of slangs and sarcasm or no dictionary preferred words. So Analysis of this category of text is often tough [4].

II. LITERATURE REVIEW

The Author [7], primary sentiment analysis especially on Twitter data. It is one of the binary classification. They used the classification features such as unigrams, bigrams, an aggregate of both and parts-of-speech tags. They examine kind of classifiers like Naive Bayes, Maximum Entropy and Support Vector Machine (SVM). By using the SVM with simplest unigram feature, they had supplied end result having 82.9% accuracy [7].

The Author [8] computes the posterior chance in Naive Bayes models using Part of Speech tags. They find SVM and CRF file, a first-class end result as an alternative unpopular measure. They have confirm by making the observations that classification overall performance will step-up with extra training data [8].

The Author [9] used two-step classifier. The first step, tweets are classified as subjective or objective. The second one group holds Twitter-specific features inclusive of retweets, hash tags, emoticons etc. They get the exceptional outcomes with the use of a SVM classifier for both steps and 81.9% accuracy for the subjectivity detection step, and the polarity detection offers 81.3% accuracy, and report a unigram baseline of 72.4% and 79.1%, respectively. For the polarity detection they discover that the meta-features are needed most and the tweet syntax features are greater important for subjectivity detection [9].

The Author [10] focuses on calculating the impact of the shortness of tweets on sentiment analysis. They accumulate tweets for five classes as entertainment, services and products, sport, companies and modern-day affairs. They document their highest quality result for binary positive/negative classification having 74.85% accuracy and 61.3% for the ternary case, both the use of Naive Bayes and unigrams [10]. The Author [11] had focuses on the challenges of the huge length of Twitter data streams. A new kappa-based, totally sliding window measure they suggest for locating category overall performance in data streams. They experiment the use of emoticons with the Stanford Twitter Sentiment dataset and the Edinburgh Twitter Corpus of [14]. They use unigrams as features. Take a look at set of the primary corpus the usage of Naive Bayes the writer data 82.45% accuracy and on the second corpus using stochastic gradient descent (SGD) the writer file 86.26% accuracy as excellent effects for the binary type project [11].
The Author [12] uses the hash tags and emoticons as noisy labels to label the data set of [13]. They used words, punctuation, n-grams, tweet period, query marks, numbers of exclamation marks, and repeat as well as capitalized words within the sentence as features. Furthermore, they comprehend the unique patterns of excessive-frequency words and content phrases and use those as features. As excellent end result for his k- nearest neighbor like classification method an average harmonic F-score of 86.0% they have suggested for binary classification [12].

III. SENTIMENT ANALYSIS

Classification of Text Mining is Sentiment analysis, which mention with the procedure of retrieving related facts and nontrivial patterns from unstructured script concept. Sentiment classification is the binary polarity type which deals with extraordinarily small number of training [5]. Sentiment class is straightforward assignment as compared to text auto categorization.

A. Data collection

We use Twitter data in our test for development and training, we use the Hash Tag data set from Twitter API. We have extracted tweets in English. We collected 7000 tweets and so on of Hash Tag [15]. The following Table I shows the number of Twitter messages and the distribution across classes.

Table I: Dataset Details

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

B. Pre-processing

Tokenization is used for splitting textual content up into words, symbols and other meaningful factors referred to as “Tokens”. Tokens may be separated by using the use of whitespace characters. The normalization process is identifying words of abbreviations available in the tweet find out after which abbreviations are replaced by Full meaning replace e.g., “OMG" by way of "Oh My God” [16]. If the word is repeated form these word will be removed into exact meaning. Delete also the Http links and Slang word like @, RT etc. and Stop words. Creating splitting the word into tokens that tokens in Unigram form. If creating Unigram word then remove stop words into Unigram word list. This word to evaluate Chi-squared. These Chi-squared passed through the classifier. These getting Unigram Word of List. Same as Bigram and Trigram to evaluate Chi-squared result. These getting Bigram and Trigram word of list. The MPQA subjectivity lexicon as word list that is labelled with sentiment polarity. We distributed lexicon features content of wordlist from the lexicon that can be represented by positive, neutral and negative

IV. SENTIMENT CLASSIFICATION TECHNIQUES

A. SVM Classifier:

SVM Classifier uses large margin for classification. It separates the tweets using a hyper plane. SVM uses the a distinctions function defined as,

$$g(F) = w\ T\ \varphi(F) + b \qquad (1)$$

’F’ is the feature vector, ’w’ is the weights vector and ’b’ is the bias vector. φ() is the nonlinear mapping from input space to high dimensional feature space. ’w’ and ’b’ are learned automatically on the training set. Here we used a linear kernel for classification. It maintains a wide gap between two classes [1].

B. Maximum Entropy

Maximum entropy maximizes the entropy describe on the conditional probability distribution. It even handles overlap feature and is equal as logistic regression which finds distribution over classes. It also follows positive characteristic exception constraints [2]

$$P_{ME}\ (c|d) = \frac{\exp \sum_i \lambda_i f_i\ (c,d)]}{\sum_{c'} \exp[\sum_i \lambda_i f_i\ (c,d)]} \qquad (2)$$

Where, c is the class, d is the tweet message. The weight vectorsdetermine the significance of a feature in classification.
It follows the similar processes as naïve bayes, mentioned above and presents the polarity of the sentiments [2].

V. EVALUATION

After finishing preprocessing all tweets are divided into two sets training and testing. Training set contains 70% of data and testing contains 30% of data. Then steps followed by feature selection which are unigram, bigram and trigram. We are finding unigram, bigram and trigram by using training and testing data sets separately

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

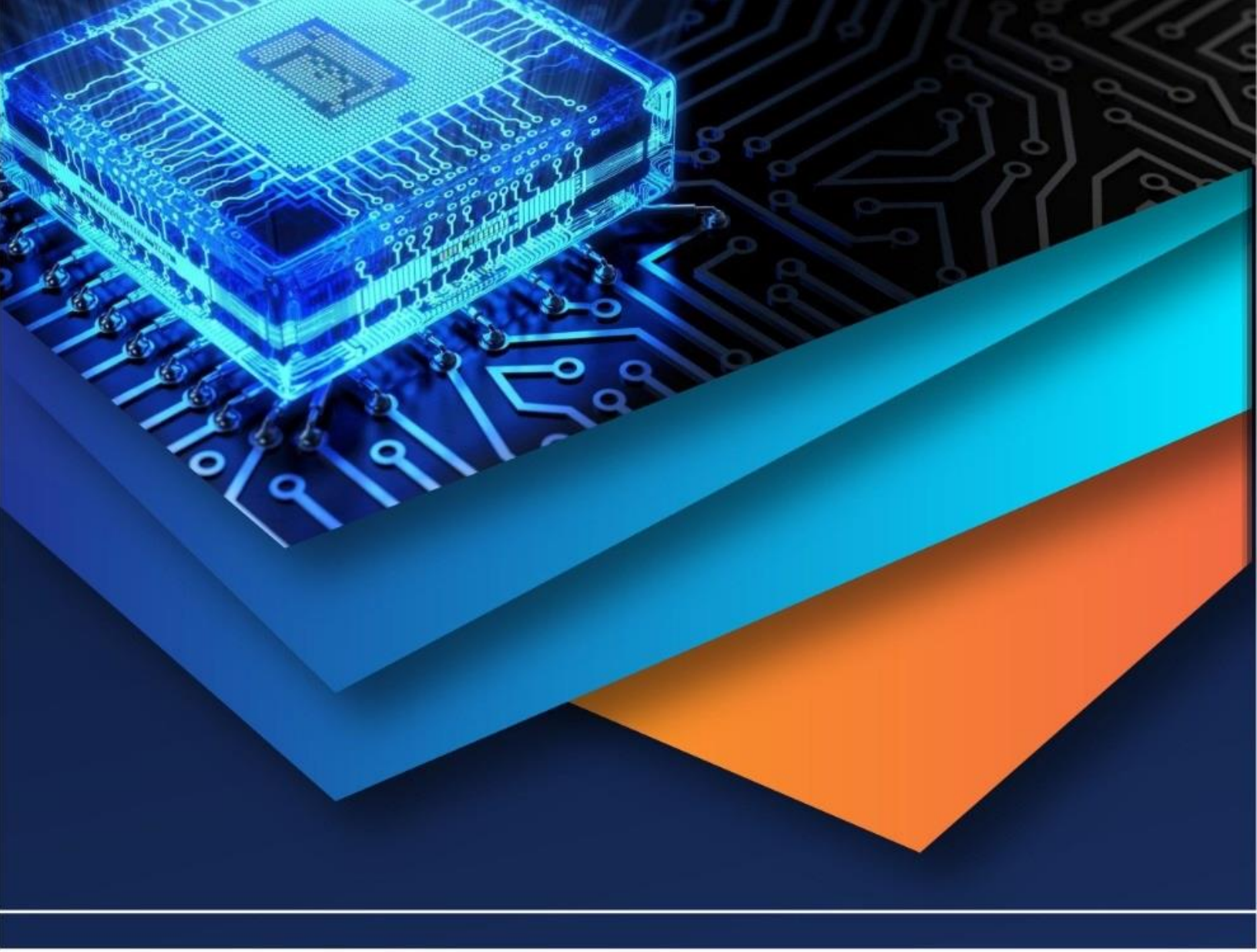
VI. CONCLUSIONS

In this paper we have demonstrated a method for automatically collect tweets using twitter API. Using that collected tweets, we preprocess and features selection for clean the tweets i.e. to remove unnecessary data from tweets. From that tweets, we use some of tweets to train the sentiment classifier. Classifier finds the positive, negative and neutral sentiments from tweets. The classifier is based on Machine learning algorithms like SVM classifier and MaxEnt classifier that use Unigram, Bigram and Trigram as their classification feature. By using Unigram, Bigram and Trigram feature selection along with two classifiers, system got such saturation of increasing training data to evaluate SVM accuracy of mean of Unigram 63.31%, Bigram 91.02% and Trigram 96.77% results. Evaluate MaxEnt accuracy of 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, Issue4, 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 acorpus 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 InternationalConference 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY