



Multi-Layer Classification and Segmentation with Opinion Mining Learning Ensemble

Ravi Kumar Kuchipudi, V.V. Jaya Rama Krishnaiah

Abstract: Social media has become the main area for advertising, events, campaigns, protests etc. in recent years. It offers the public a forum for expressing their opinions and beliefs to the masses. The user beliefs, habits and interests of businesses are extremely important and give users a glimpse into their thinking. Data mining is one of the tools that enables these businesses to extract useful information from user data that can be examined to generate a set of knowledge and identify a user opinion that helps companies to create user-specific products. Twitter Data Mining and other social platforms are very important since its enormous user base consists of a mixture of thoughts and opinions that may be used to anticipate results of campaigns, product assessments and similarity if correctly studied. This research provides a classification system to perform Twitter Opinion Mining Segmentation based on Ensemble Learning. The suggested approach can detect and filter out boxes and uses text segmentation for efficient text classification and voice tagging.

Keywords: Machine Learning, Supervised Learning, Text Analysis, Sentiment Analysis, Natural Language Processing.

I. INTRODUCTION

Twitter is a micrologging website that gives users the opportunity to express their views in brief, public tweets. Twitter has around 319 million active users since 2006, making it a source of information. People share their feelings on problems, campaigns and products which could be utilised to reduce the general likelihood of this problem, campaign or product. An architecture has been suggested for a Twitter opinion miner to build an opinion miner for Twitter. The aforementioned miner uses segmentation techniques to effectively mark the utterance of phrases in natural language detection[1]. Segmenting material into meaningful phrases and tagging them accordingly to accurately quantify the extent of meaning the phrase transmits. As twitter is multi-dimensional, one classification is just not sufficient to appropriately classify this data. The architecture therefore leverages ensemble learning to address this issue. Extensive pre-processing is necessary before the opinion rules are extracted from the data. Spell checks, bot sensors, sarcasm detectors, lemmatizers and tokenizers are included in the pre-processors.

Manuscript received on October 18, 2021.

Revised Manuscript received on October 27, 2021.

Manuscript published on October 30, 2021.

*Correspondence Author

Ravi Kumar Kuchipudi*, Research Scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur (Andhra Pradesh), India E-mail ravikumary4@gmail.com

Dr. V.V. Jaya Rama Krishnaiah, Associate Professor, ASN Engineering College, Tenali (Andhra Pradesh), India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

The following information suggests a humble Twitter opinion mining architecture learning. Liu et al. [2] define opinion as quintuple, (ei, aij, ooi, h, tl), where ei is the entity name, aij is the ei aspect, ooi is the aij of the ei aspect, h is the bearer of the opinion and tl, when the h opinion is conveyed. [2] Liu et al.. Ooi, h, tl may be positive, negative or neutral, or may communicate an opinion at varied strength/intensity levels. If an opinion is about the entire entity itself, the special aspect GENERAL is used to indicate it. The use of natural language processing, text analysis and computational linguistic to reduce these favorable, negative or neutral thoughts is called Opinion Mining. Twitter includes multi-dimensional data and is one of Micro logging's most popular platforms. People offer their comments on items, problems and twitter campaigns for the millions to read. The views of mining people on that product or campaign can target a product or campaign that gives them an insight into the general agreement of the people with regard to that product or campaign. A corpus of textual tweets was developed to train the classifiers to classify the broad emotions of the subject matter appropriately. The corpus was then rigorously pre-processed to make it fit for mining. Then the processed corpus, Naïve Bayes Classifier[11]. Senti Word Net[12],[13] analysis and Emoticon analyses. The individual results of these students have been aggregated to create an opinion which gives an indication whether a tweet is positive or negative..

II. RELATED WORKS

Extensive literature surveys were carried out and the following information was acquired from the survey.

- A hybrid categorization is introduced to [3] the Twitter opinion mining architecture. The study analyses the many obstacles of Twitter feed sentiment analysis and provides a technique for real time text mining and sentiment analysis, utilising a three-way classification, by examining the intensity of feelings. It includes data pretreatment utilising techniques like slang/abbreviation analyses, lemmatization, rectification and stopword removal and also tests the accuracy of sentiment recognition on 6 twitter data sets of the proposed algorithm. In conjunction with these preprocessing techniques, the architecture proposed includes real-time tweet access via a Twitter streaming API and a three-stage categorization to determine the tweet feeling. The first level is an Emoticon classifier, which uses a sample set to detect and categorise the emoticons as positive or negative. The second level is a polarity classifier that uses "sacks of words" The Senti Word Net Classifier is followed.



Each of these classifications assigns the tweet a score that determines the polarity of the tweet in relation to the subject. A tweet with a score of 0 is called neutral considering all 3 classification techniques. The results are shown graphically, which compare the outcomes based on confusion matrices, accuracy, precision, F-measurement and recall.

- Taking account of the critical role played by the media and news in forming an individual's own attitude on any subject[4], Senti Word Net introduces a news headline opinion mining approach. It employs two algorithms to accomplish a given function for each algorithm. Words are pre-processed using Python's libraries and a part-of-speech mark is assigned to each news headline word to establish the headline's feeling. The POS-Tagger is used to identify which word is what component of this algorithm is spoken and lemmatized. The second algorithm translates the words into Senti Word Net, which further gives each word a positive or negative value and detects the overall polarity of the headline to characterise its influence or opinion on a given theme. The above strategy was used in a series of 500 headlines spread over 30 days and the average difference per day was 2.7. This illustrating the efficiency of the opinion mining on data supplied by offline users such as headlines for newspapers, instead of data generated by the internet user, such as blogs, reviews, tweets, etc.

- Comparative studies of existing opinion mining techniques, including machine learning and lexicon-based approaches[5], cross-domain and cross-lingual methods and some evaluation metrics were offered. The research demonstrated that machine learning methods such as SVM and Naive Bayes were of the highest accuracy and may be considered basic ways of learning, but in some circumstances, the approaches based on lexical substances were effective in the human text. The Bigram model gave better outcomes than the Unigram and Trigram models among the numerous models evaluated and has the capacity for adaptation to a number of areas and languages. The machine learning methods identified the subject in question as a classification issue and the SVM and Naive Bayes methods were utilised as supervised research methods. The Lexicon-based approaches use dictionary and corpus-based analytical methodologies. The whole task of sentiment analysis was divided into three tasks, classification of subjectivity, classification of sentiments and other supporting tasks. Those tasks were then carried out in multiple grain levels, such as word-level, sentence-level, document-level and aspect-level. The dataset used for the investigation was Stanford's publicly available Twitter dataset. The problem of sarcasm identification is a huge obstacle, i.e. a phrase which contains solely good words yet implies a negative feeling.

- A system for recognised and analysed the feelings of restaurant and laptop review using Support Vector Machines (SVM) was proposed in [6]. Another text classification technique was proposed to identify the review phrases component. Dates from SemEval 2014, SemEval 2015 and SemEval 2016 were acquired. The reviews were categorised into many dimensions such as ambience, food, prices etc. The overall task consisted of the following subtasks: feeling term extraction, determination of polarity, aspect term extraction and aspect word categorization. The algorithm initially detects the term in the sentence and then categorises it adequately, and then identifies and classifies the term in

the sentence. The experimental evaluations anticipate the usefulness of SVM in text categorization in comparison with other methods of classification. In most cases the accuracy achieved using SVM for categorization of emotion and aspect was better compared to other classification approaches such as KNN and Naive Bayes.

- Opinion mining provides user thoughts and reviews at three document, sentence and feature levels. In [7], two strategies for extracting features were proposed. These approaches consist of four steps, the construction by lexicons for the orientation of user opinions; the preparation of tokenisation and syntactic parsing of addition; extracting features using frequency extraction features and extracting grammar features, and finally determining polarity. In the end stage, every review feature produced a record including five features, including feature, polarity, date, author and type. The polarity was determined by adding the polarity of opinions describing the characteristic and taking into consideration the function of negatives in the statement. The dataset utilised was created from mobile phone reviews obtained by users from the digikala website and the Virastyar software conducted the preprocessing. The findings of the performance of the two proposed approaches demonstrated that the dependence grammatical method is better at extracting features than the frequency-based method. However, the polarity of freshly collected opinions is not well determined. Tang et al [8] proposed a methodology of combined segmentation and classification for sentiment analysis in order to efficiently handle the inconsistent feeling polarity between a sentence and the words it contains. A Joint Framework for segmentation and classification (JSC) was developed which at the same time performs sentence segmentation and sentence level classification of feelings. The JSC framework included the Segmentation Model and the Classification Model in two models. The model of segmentation was divided into two parts: the model generating the segmentation candidate and the segmentation ranking model to efficiently create segmentation candidates for each sentence using a phrase table beam search caused by massively large companies. The technique to the identification of these sentences was based on the frequency of unigrams and bigrams in the corporate environment. The concatenation of both features was aimed to provide two kinds of characteristics for the sentence segmentation, phrase-embedding and segmentation specific feature, and the final feature representation for each segmentation. In addition, each segmentation candidate was allocated a scalar in the ranking model to represent the utility of the result for the sentiment categorization. The partition score was estimated with a log-linear model. A supervised learning system was implemented for sentiment classification from sentences with manually labelled feeling polarity. For a test sentence, its candidates were generated for segmentation and the score was determined for each segment. The top K candidates were selected and the final outcome was voted on their anticipated feeling polarity from the classifier.

III.METHODOLOGY

A mechanism was discussed [9] that allows data from the popular micro-blogging site Twitter to be used and interpreted to determine public attitudes. The analysis of the Apple iPhone 6 tweets and features contained specific popularities. The approach for discovering feelings and opinions was local, i.e. the popularity of the particular product was measured in numerous places. Data has been obtained using the public Twitter API, and pre-processing has been performed to filter out superfluous messages. An AI technology has been used to determine the user's gender because Twitter doesn't ask for sex while building a new account. Different national and city-wide indicators were discussed, together with gender-separate metrics for various cities in the USA. SentiWordNet was used to evaluate the feeling of a tweet and provide it with a numerical score ranging from -1 to +1. In order to detect trends and variations in feelings, multiple comparisons have been done on various levels. Initial comparisons were made at an international level and a more extensive examination of user cities and genders was carried out. The deductions made were that very negative feelings were recorded to the iPhone 6 screen and touch due to the bending problem surrounding the phone during its initial release. On the other side, the phone's camera was highly positive and was applauded by both regular users and reviewers.

IV.PROBLEM STATEMENT

After consultation, the following challenges were discovered in the rich literature on opinion mining, text classification and machine learning.

Why One algorithm is just not sufficient to classify and view multidimensional data from Twitter.

- Social media data are prone to orthographing errors, sarcasm and Internet lingo affecting classification quality.
- Learning algorithms are efficient and depend on the data set. No single method can work successfully on any data collection.
- Scalability in opinion mining is a huge problem. Increasing the data set size can affect the classification's accuracy.
- Twitter is full of bots that take advantage of machine learning algorithms. This affects the accuracy of learning algorithms.

V.PROPOSED ARCHITECTURE

The Opinion Miner was split into 3 sections, data collecting, preprocessing and categorization and assessment. You may find the architecture of each module below..

4.1 Proposed Miner Architecture

As illustrated in Fig. 1, Twitter streaming API tweets were collected on the basis of a specific input query. English tweets have been filtered and tweets stored in JSON format have been created. However, this corpus requires pre-processing for precise mining. Twitter Bot or Not[10] have been subject to the tweets to filter out bot accounts from the corpus. Twitter bots influence learning quality and often degrade the accuracy of the learning algorithm. URLs, identities and hashtags have been removed after the bots have been filtered. Spell checking has been carried out to

correct typographical errors and incorrect terminology. Emoticon classification was then utilised to extract tweets from emoticons. Lemmatization[15] and stemming have been done on every tweet word for part of language tagging after which classification features have been extracted. Each tweet that ranked the top k portions has been segmented. These top K segments have been used as classification features for the Naïve Bayes classification and tweet. Applied to each tweet were the Emoticon classifier and the Senti Word Net classifier. Each tweet was marked by a combined weighted score from each of the three classifiers (CWS). The procedure is shown in the following algorithm and in Fig. 2.

4.2 Combined Weighted Score Calculation Algorithm (CWS)

Algorithm

Input: String query

Output: Predict tweet-based opinion and feeling.

1. Begin
2. Enter a string query.
3. Retrieve input query data from the Twitter API.
4. Filter out tweets in non-English languages.
5. Filter out bots tweets.
6. Remove tweets duplicate.
7. Save JSON tweets.
7. Every tweet,
 - 8.1 Pre-process procedure (tweet)
 - Delete URL.
 - Remove your username.
 - Removal of hashtags.
 - Emoticon Classifier Emoticon Extract.
 - Replace abbreviations and slangs.
 - Spell Application Verification and adjustments.
 - Lemmatization.
 - POS tagging and stemming.
 - Classification extract characteristics.
 - 8.2 End of proceedings
 - 8.3 Segmentation of the procedure (processed tweet)
 - Generate candidate segmentation.
 - Calculate the segmentation score based on the segment and the sentence table
 - Choose the top segments K-ranked.
 - 8.4 Final procedure End
 - 8.5 Classification of procedure (processed tweet)
 - Classify Emoticon Classifier tweet.
 - Classify tweet with the Classification of Naive-Bayes utilising segments as features.
 - Use SentiWordNet to classify the tweet.
 - Combine the result acquired from the three students.
 - Save the categorization prediction result.
9. Finish For
10. End

Combined Weighted Score Calculation Formula (CWS)

$$CWS = 0.5*N_b + 0.3*S_w + 0.2*E_c$$

Here,

N_b: Naïve Bayes score of the tweets

S_w: Score from SentiWordNet classifier

E_c: Score from Emoticon classifier

VI. EXPERIMENT SETUP

The proposed architecture was multiple times tested for various corpus subsets. The result in the following parts is an average of five independent runs on a 5000-sample dataset with 90/10 fractions, i.e. 90% of the samples for training and 10% for examination. The process took 1012 seconds on average.

5.1 Segment Selection and feature Extraction

The Brown Corpus [14], which has 3,583 and 3,585 entries, was used to produce a number of bigrams and trigrams. Table 1 and Table 2 offer partial lists. Complete lists with the author are available. If a segment exists in these sets, for each tweet, it was only chosen otherwise. The segments picked were those of sentimental importance analysis. Top k segments were extracted from the corpus of processed tweets and these were used as features for Naïve Bayes classification.

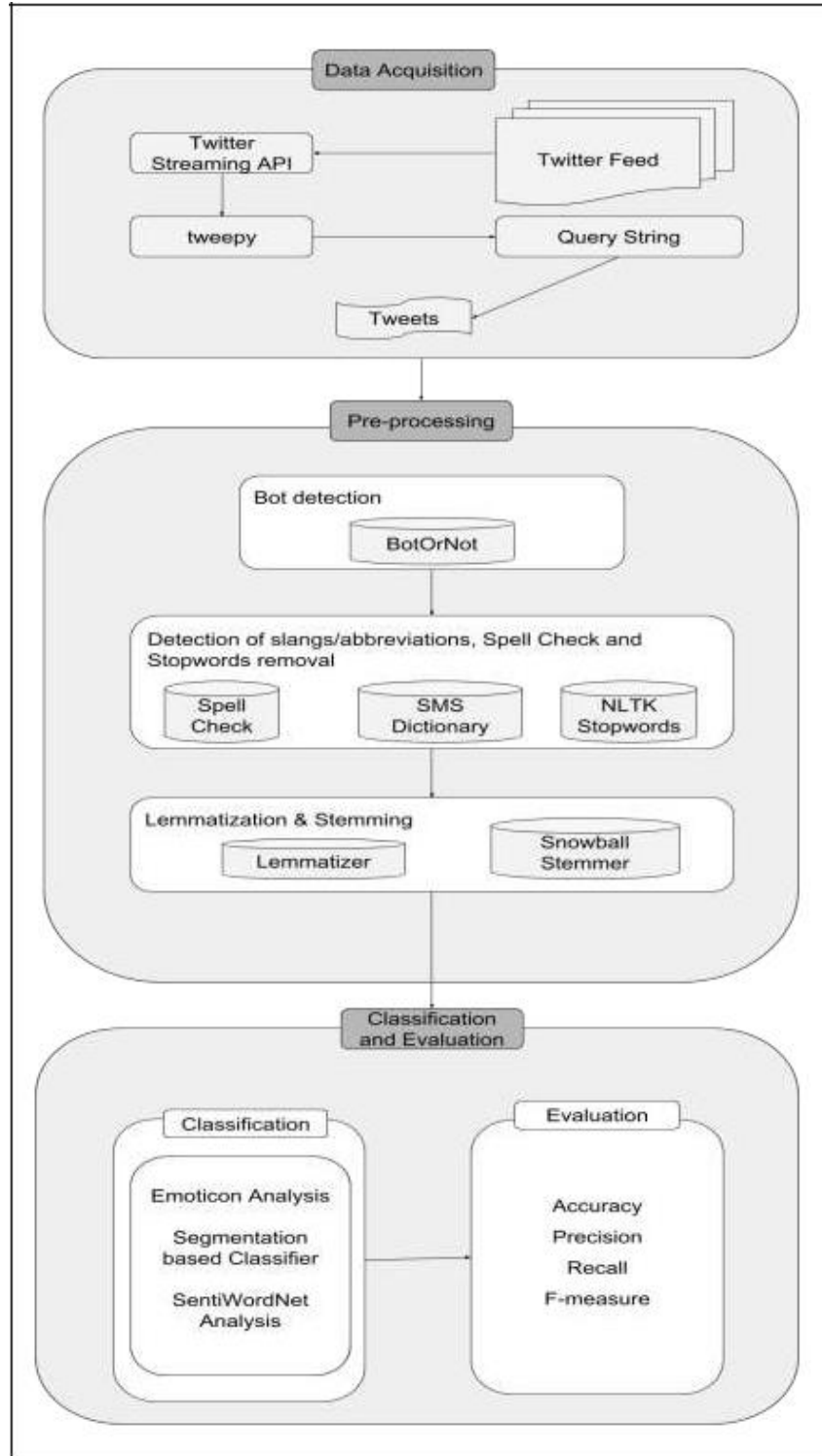


Fig. 1: Twitter Opinion Miner Architecture

Fig. 3 displays the list of top k segments derived from tweets based on their probability in Naïve Bayes. The first column lists the extracted features. The second column lists the relevant feeling score. 0:1 is a negative feeling whereas 1:0 is a nice feeling. The third column represents the corresponding probability of the segment in Naïve Bayes. These segments are the features of the classification.

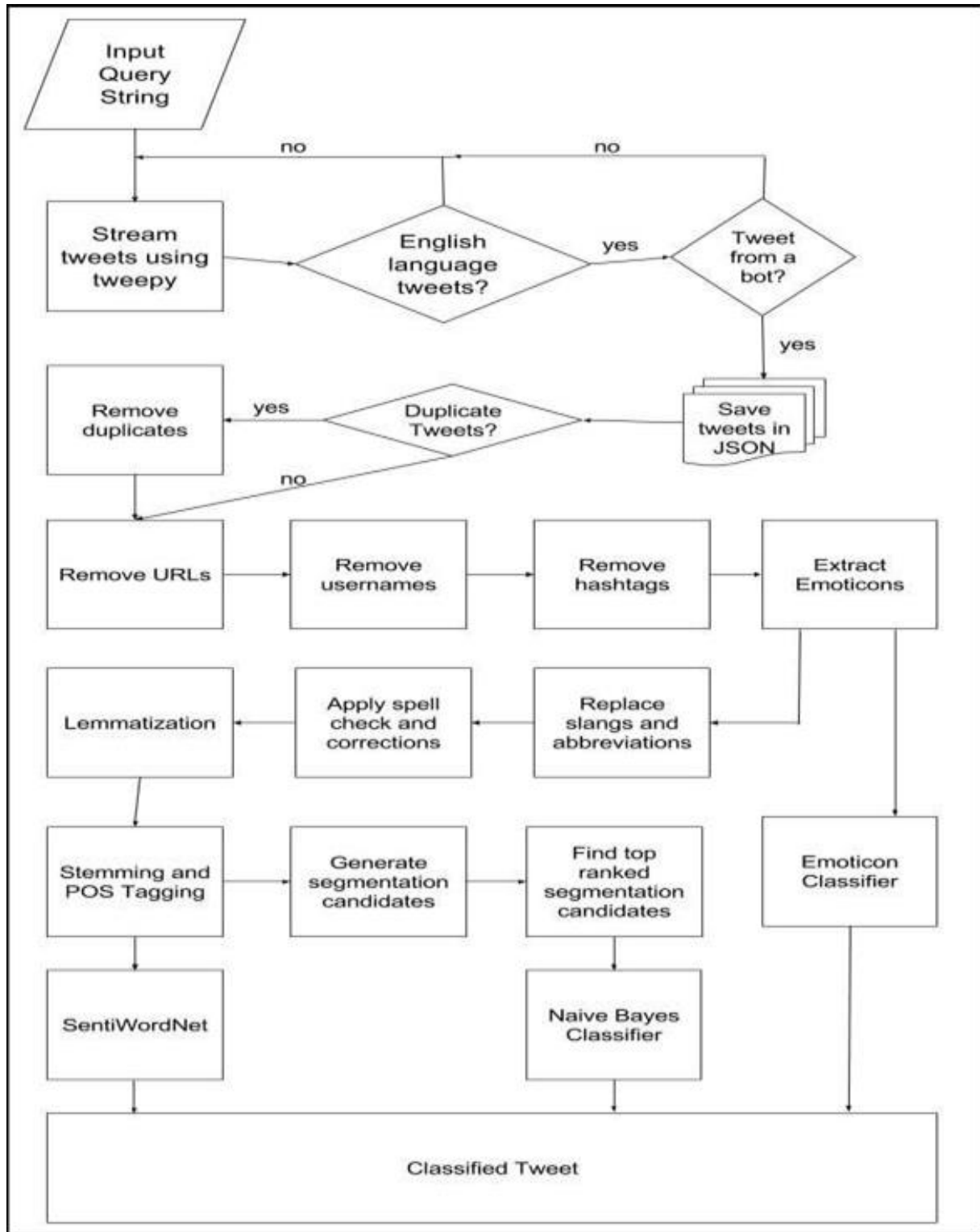


Fig. 2: Flow chart depicting Opinion Mining of Tweets

5.2 Dataset for classification

For sentiment analysis, a data collection of 1,578,627 categorised tweets was employed. Each row is marked 1 or 0, correspondingly for positive and negative feeling. For training and 10 percent for testing, 90 percent of sample size was employed. The remaining data were dumped into a

JSON file after initial filtering for further processing and classification. After processing, the positive and negative tweets are shown in Fig. 4..

Table 1: Partial list of bigrams

Bigrams
United States
New York
I don't per
cent
years ago
I think I
could I
said AfAf
I know I
would
said I
He said
The first
could see
Rhode Island
Of course
In addition
I thought
Ididn't
But I
last year
first time
It would
don't know
For example
I told

Table 2: Partial list of trigrams

Trigrams
I don't know
World War 2
United States America
New York City
Government United States
I could see
I said I
New York Times
The New York
I don't think
The United States
basic wage rate
A. Notte Jr.
Drug's chemical name
John A. Notte
New York Central
I would like
per capita income
I told I
World War 1
words electronic switches
I know I
index word electronic
index words electronic
small business concerns
word electronic switch
I don't want
result obtained item
two years ago
uniform fiscal year
Department Economic

Affairs

```
Most Informative Features
contains(sad) = True          0 : 1 = 20.7 : 1.0
contains(excited) = True     1 : 0 = 13.2 : 1.0
contains(enjoy) = True       1 : 0 = 12.2 : 1.0
contains(especially) = True  1 : 0 = 10.6 : 1.0
contains(love) = True        1 : 0 = 10.6 : 1.0
contains(wonderful) = True   1 : 0 = 10.6 : 1.0
contains(remember) = True    1 : 0 = 10.6 : 1.0
contains(bad) = True         0 : 1 = 10.0 : 1.0
contains(thx) = True         1 : 0 = 9.0 : 1.0
contains(getting ready) = True 1 : 0 = 9.0 : 1.0
None
training time: 1881.4409999 s
testing time: 373.823000193 s
{'accuracy': 0.8579999999999997}
```

Fig. 3: List of most informative features

```
Positive tweet:
Tweet: Jesus Gives Me Hope That One Day I will Rise to Heaven Too. Happy Easter
to Those Who Fears the Lord When No One Is Looking. #jyc
Processed Tweet: Jesus gives hope one day rise heaven happy easter fears lord one looking
Negative Tweets:
Tweet: Feeling very #unmotivated and kinda #disappointed
Processed Tweet: feeling very unmotivated kinda disappointed
```

Fig. 4: Original and Processed tweets

5.3 Emoticon Classifier

The emoticon classifier identifies positive and negative emoticons in accordance with the regular expressions in Appendix A. The feelings are properly assigned. Fig. 5 displays the emoticon classifier output.

```
Tweet:
Great win tonight @thedujarations! Special mention to @aj191 for an amazing deb
ut and @Raina on playing his 150th :) #IPJ
Sentiment: 1
Prediction: 1
Tweet:
I miss how happy I used to be... :(
Sentiment: -1
Prediction: -1
```

Fig. 5: Emoticon Classifier Output

5.4 Senti WordNet Classifier

Classifier SentiWordNet utilises SentiWordNet 3 to assign an emotion score to each tweet. A positive mark means a positive and a negative mark means a negative mark. Fig. 6 shows the scores for sampling tweets using the classifier SentiWordNet..

```
Tweet:
Fantastic. The real way to start a swachhagraha. Start young. All our schools sh
ould follow this example.
Sentiment: 1
Prediction: 1
Tweet:
Another day another Muslim terrorist attack & people want to import these savage
s to America. #sad
Sentiment: -1
Prediction: -1
```

Fig 6: Senti Wordnet classifier output

5.5 Results

Table 3 below shows a comparison examination of the outcomes of the suggested miner and the miner proposed by Bashir et al. The accuracy, reminder and f-measurement measures were determined and compared below.

Precision is how often a feeling rating was correct. The ratio of true positive against all positive ones is mathematically (true or false).



Reminder is the ratio between classified positive properly by the classifier and the positives manually classified (true positive + false negatives). Mathematically,

Table 3: Comparison of proposed miner with existing one

Property	Proposed	Bashir et. al.
Accuracy	85.80%	85.80%
Precision	84.80%	86.00%
Recall	85.20%	82.20%
F-Measure	84.98%	84.05%

VII. COMPARISON SIMILAR OPINION MINERS

The Opinion Miner suggested has the following benefits over similar opinion miners.

- Scalability: The suggested miner is extremely scalable and can handle big datasets with moderate hardware. The whole process took 1012 seconds for a sample of 5000 tweets.
- Bot Detection: before mining, the suggested architecture will filter out postings from twitter bots. This is done with the BotOrNot algorithm. The eradication of these bots improves the precision of the mining process.
- Context spell check: The architecture uses contextual spell checks to fix words, taking into account their context and also identifying prominent English idioms and phrases.

VIII.CONCLUSION

The article deals with Twitter opinion mining and sentiment analysis. It identifies several significant difficulties in social media opinion mining, such as orthographic errors, and affects mining quality. Section 4 proposes a humble architecture that overcomes these difficulties. Proper measurements have been obtained in the same and separate data sets listed in Section 5 after numerous independent runs. The essay compares similar current approaches. The proposed miner achieved a precision of 86 percent and a recall of 84 percent. The f-measurement was estimated to be 85.98%.

FUTURE SCOPE

The suggested architecture works in English solely for tweets and filters tweets in other languages. The future study will focus on the development of an architecture capable of handling several languages. Monitored learning algorithms and SVMs can be used to improve the precision further.

REFERENCES

1. M. Bayomi, K. Levacher, M. R. Ghorab and S. Lawless, "OntoSeg: A Novel Approach to Text Segmentation Using Ontological Similarity," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, pp. 1274-1283, 2015. DOI: 10.1109/ICDMW.2015.6
2. B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis", Mining Text Data, pp. 415-463, 2012.DOI: 10.1007/978-1-4614-3223-4_13.
3. F. Khan, S. Bashir and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme", Decision Support Systems, vol. 57, pp. 245-257, 2014. DOI: 10.1016/j.dss.2013.09.004

4. A. Agarwal, V. Sharma, G. Sikka and R. Dhir, "Opinion mining of news headlines using Senti WordNet", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016. DOI: 10.1109/CDAN.2016.7570949.
5. V. A. and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications, vol. 139, no. 11, pp. 5-15, 2016. DOI:10.5120/ijca2016908625.
6. D. K. Kirange and R. R. Deshmukh, "Emotion Classification of Restaurant and Laptop Review Dataset: Semeval 2014 Task International Journal of Computer Applications, vol. 113, no. 6, pp. 17-20, 2015. DOI: 10.5120/19829-1680W
7. E. Golpar-Rabooki, S. Zarghamifar and Jalal Rezaeianour "Feature extraction in opinion mining through Persian reviews", Journal of Artificial Intelligence and Data Mining, vol. 3, no. 2, 2015. DOI: 10.5829/idosi.jaidm.2015.03.02.06
8. Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu and Ming Zhou, "A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1750-1761, 2015.DOI: 10.1109/TASLP.2015.2449071.
9. S. Anwar Hridoy, M. Ekram, M. Islam, F. Ahmed and R. Rahman, "Localized twitter opinion mining using sentiment analysis", Decision Analytics, vol. 2, no. 1, 2015. DOI: 10.1186/s40165-015-0016-4.
10. C. Davis, O. Varol, E. Ferrara, A. Flammini and F. Menczer, "BotOrNot", Proceedings of the 25th International Conference Companion on World Wide Web- WWW '16 Companion, 2016. DOI: 10.1145/2872518.2889302.
11. G. Webb, J. Boughton and Z. Wang, "Not So Naive Bayes: Aggregating One-Dependence Estimators", Machine Learning, vol. 58, no.1, pp. 5-24, 2005 .DOI: 10.1007/s10994-005-4258-6.
12. S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in LREC, 2010. ISBN :2-9517408-6-7.
13. A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining." 2006.
14. "Brown Corpus", Essex.ac.uk, 2017. [Online] Available:http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html. [Accessed: 16- Apr- 2017].
15. A. Uysal and S. Gunal, "The impact of preprocessing on text classification", Information Processing & Management, vol. 50, no. 1, pp. 104-112, 2014.DOI: 10.1016/j.ipm.2013.08.006

AUTHORS PROFILE



Ravi Kumar Kuchipudi, Research Scholar, anu, Computerscience and Engineering., Having 10 years of Teaching Experience in Various Computer Areas, Published Various Papers in Reputed Journal



Dr. V.V. Jaya Rama Krishnaiah, Associate professor, ASN Engineering college ,Tenali , he has 24 years of experience in teaching and research and has published 30 research papers in reputed journal of scopus and UGC