MANTECH Publications

Journal of Research in Computer Science and Engineering
Volume 7 Issue 2
ISSN- 2457-0818 (Online)

An Analysis of Python for Large Amounts of Data and Sentiment Analysis

Pramod Kumar Mundra[1], Savita Kumbhar[2]

Assistant Professor[1], Student[2]

Department of CSE

AISSMS College of Engineering

Corresponding Author's Email: -*kumbhar19sept@gmail.com*

Abstract

"At the moment, social media has a wealth of information on its members. Some of the methods for extracting information from social media provide us different applications in various disciplines and investigations. Extraction of information from social sites/media in Product Analysis provides a lot of benefits such as knowledge about the newest technology, real-time market scenario updates, and so on. Twitter is one of the social media platforms that allows users to submit tweets with a restricted amount of characters and share the message (tweet) with their followers. It enables developers to access the data for their purposes. Details are gathered and sentiment analysis is conducted in the implemented module. Tips and information can be offered to the user based on the outcomes of data and emotional analysis. The running module may do data and sentiment analysis on data from numerous sectors, as well as user views and suggestions on various items. These outcomes can help businesses stay current, for example. The built system may assist estimate the consequences of various items and actions in numerous domains using this technique."

Keywords: - *Social Media, Numerous Domains, Python, Methods*

INTRODUCTION

In today's world, social media and numerous accompanying applications allow all users to voice their ideas on a certain issue and demonstrate their attitudes by like or disliking information.

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

All are constantly gathering social media acts and creating great variety, volume, velocity, value, and variability data referred to as big social data. This type of data refers to a vast number of individual ideas that may be analysed to understand people's digital inclinations. Several scholars have expressed an interest in utilising massive amounts of social data to explain, identify, and forecast human mindset in a variety of disciplines. This type of processing entails a variety of study approaches, including text analysis. In reality, 85% of web data is text, and text data analysis has become a critical component for determining public mood and valued opinions on content. Sentiment analysis, also known as opinion mining, seeks to discover individuals' feelings on an issue by analysing their social media postings and various forms of behaviours. The polarity will then be divided into three groups, such as positive, negative, and so on.

Sentiment analysis can be divided into two categories:

- Lexicon analysis, which aims to find or calculate the polarity of a document from the semantic analysis of words or phrases in the document. However, applications which are based on lexicon analysis never consider the studied context.
- Machine learning (ML), it involves formation of models from labeled training dataset (instances of texts or sentences) in order to determine the orientation of words and phrases in a document. Studies that used machine learning methods have been carried out on an important topic.

These two analysis methodologies have been widely utilised on big data to collect public criticism in order to gauge internaut satisfaction with a subject (services, goods, events, themes, or different people) in many fields such as health, politics, and marketing. However, the results can change with a decent degree of precision at times and not at others. The failure is usually caused by the difficulties of opinion mining, such as the semantic analysis of a term whose meaning relies on context. Our goal in this work is to tackle semantic analysis by proposing an efficient and new adaptive technique that relies on social media postings and big data architecture to assess internauts' sentiments and behaviour toward a certain issue in real-time. The suggested method is divided into three steps, as indicated in Fig. below.

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

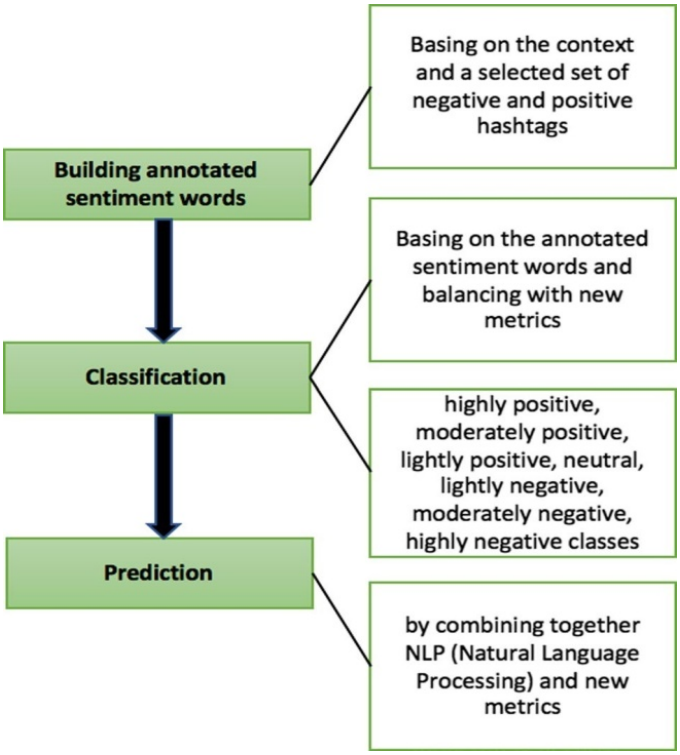


Figure:-1 Stages of the model

RELATED WORK

Several studies have focused on social media analyses, particularly when they are tied to major events that would get a lot of attention, such as presidential elections. Social media has become an essential tool for both politicians to communicate directly with voters and discuss their programmes, as well as voters to voice their opinions on each candidate. This heavy use of social media platforms has sparked widespread interest in academic study, and several sorts of contributions have been made to track such occurrences, which include millions of internet or social media users.

These contributions could be categorized into four categories:

1. Opinion-based approach in which opinion mining methods are used by the authors for their models, detailed in "Theoretical basis" section, in order to classify posts related to a candidate. There are two categories classes:

- Sentiment classes, in which sentiments are, classified under two classes (positive and negative). Other authors have added a third class (neutral).

- Context classes, in which polarity depends on the context. Conover et al. have classified the posts in three classes which are Left, Ambiguous and

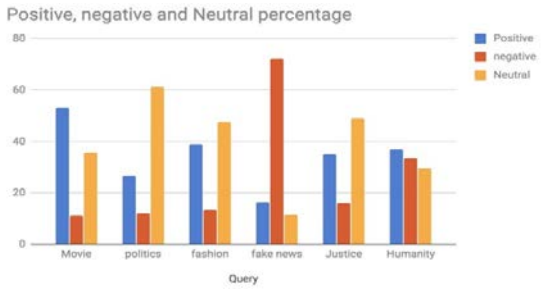
Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Right while authors in have based their prediction model on two classes (Pros and Anti) for each party.

2. Volume-based approach, in this researchers aim is to determine the candidate who is elected based on the number of tweets which mentioning them (% mention) or retweet volume. In fact, researchers in have discovered an interesting correlation between name mentions percentage or retweet volume and the vote.

- Opinion and volume (OV)-based approach in which both opinion mining and volume approaches are combined together. Researcher Finn et al. discovered a new approach to measure political polarization without the use of text. They have used a co-retweeted network, as well as the retweeting behavior of the social media users.

- Emoji-based approach, in this the classification of posts is based on the use of emoji. Researchers selected differente moji and categorized them into different categories such as happy, sad, fear, laughter, and angry classes, then, find the sentiment of the first emoji in the post.



Text mining

- Text mining may be defined as the process of extracting meaningful information from unstructured textual data. This procedure is divided into two primary stages: The analysis phase refers to the process of organising text via the use of linguistic analysis techniques such as recognising words, sentences, grammatical roles, and connections. It entails numerous methods:
- Language identification, it is the process of determining the natural language in which a given text is written.
- Tokenization, it is the process of segmenting a sequence of strings into words and sentences by removing some characters like punctuation marks.
- Filtering, this process consists of applying filters such as removing empty words.
- Lemmatizations in this process different inflected form of a word are grouping together by removing plurals,

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

genders and conjugations. Then, we analyze them as a single item.

- Named-entity recognition, it is the process in which we are going for searching text object which can be categorized in classes such as persons, dates and localization.

The output of the first phase is evaluated and interpreted by interpretation phase by using methods of data mining. The purpose is to find patterns, relevance, novelty, and interestingness. Text mining does not allow extracting opinion, so other techniques are also combined in it. Next, we will present some of those techniques.

Opinion mining

Opinion mining is the technique of science in which we are using text analysis to determine the sentiment analysis of a text (positive, negative or neutral). It can be determine under different terms: sentiment analysis subjectivity, analysis of stance. One of its important application is to understand and track the mood of the users of social media about a specific topic in different domains such as marketing, health and politic. These product reviews can be used by Potential buyers for making their decisions related to the product.

These are following approaches to carry out opining mining:

Lexicon based approach

Sentiment lexicon and a collection of known sentiment terms are used in lexicon-based approach. It is mainly divided into two approaches i.e. dictionary-based approach and corpus-based approach. Dictionary-based approach finds opinion words in the text and then finds semantic orientation of those words in the dictionary.

There are number of dictionaries such as Senti Word Net and they can also be created manually. Corpus-based approach is used to find opinion words in a context specific orientation. It generally starts with a list of opinion words and then other opinion words are going to be determined in a large corpus.

Learn based approach

The learn-based approach depends on the famous ML algorithms (i.e. supervised and unsupervised methods). In the supervised methods, we will train the model through a large number of labeled documents. The most famous ones for opinion mining are: Support vector machine, maximal entropy principle, and the Naive Bayesian classification. These methods having a high accuracy in determining the polarity

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

in the domain that they are trained on but their performance fall precipitously when the same model is used in a different domain. The unsupervised methods are used when it is difficult to find labeled training documents. However, because of their bad performances in this area, it is rarely used at present.

Hybrid approach

The hybrid approach uses both, the lexicon and the learn-based approaches. It uses the lexicon-based approach for determining the sentiment scoring. Then, training data for the learn-based part will represent by these scored documents. Hybrid approach is widely used because of its two qualities i.e. improved or high accuracy and due to its stability that comes from ML powerful and the lexicon based approach, respectively.

All three approaches i.e. Hybrid, learn-based and Lexicon-based approaches have been widely used in different domains, in different ways and improved their efficiency by several searchers.

By referring to opinion mining approaches, we present a method that analyzes social media posts and extracts user's opinion.

THE PROPOSED METHOD

In order to build a model for sentiment analysis, we propose three stages based methodology in this paper that contains, first building sentiment words, then classifying and balancing this set of words before executing the prediction algorithm.

The descriptions of the three stages are explained below:

Let: Yi, i=1,n be a set of products, services or persons that we will aim to compare in a specific context. Let's consider D= {Y1, Y2, ... , Yn}.

FIRST STAGE: CONSTRUCTING DICTIONARIES

Based on hashtags description various researches in social media analysis have identified whether the intention behind a post is positive, negative or neutral. However, they have used a manually defined very large set of annotated hashtags (which may take large amount of time) or they have combined these latter with dictionaries in order to improve classification posts accuracy. At the beginning stage, we use a small set of hashtags, in order to build dictionaries of words, annotated with the word's semantic orientation for a given context as following:

We are going to assume that each and every word in a tweet that contains a

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

negative hashtag is negative and a tweet that contains a positive hashtag is positive, then, we process it by different steps.

Step1 Contains posts which are related to Yi. The aim of our approach is to compare Yi, we will going to identify hashtags with high frequency as the most popular hashtags for each Yi. After that we will classify a small set of them manually into negative and positive classes and have to collect related data for each class separately. We will classify collected tweets into two classes i.e. positive or negative based on the upper defined polarity of hashtags.

Step 2 Consists of data which is preprocessing classified extracted from hashtags. Social data is an informal type of data that could contain spelling mistakes and non-textual information, hence there is need of a preprocessing step and for that we are going to apply various filters on tweets as following:

- Tokenization is a process that contains sub-step which consists of identifying nouns, verbs, adverbs, adjectives, URLs, common emoticons, phone numbers, HTML tags; Twitter mentions hashtags, and repetition of symbols and Unicode characters.
- Conversion is a process where all words will be converted to lowercase and replace more than two of the same consecutive letters in a word with only one occurrence of the letter (e.g., we replace Sunny by sunny and ANGER by anger).
- Stemming is a process where we remove plurals genders and conjugation (applying morphology stemming).
- Filtering can be defined a process where we enhance the indicators by applying other various filters and sentiment indicators such as hashtags.
- Indicators can be defined as adjectives and verbs which are good indicators for positive and negative sentiment analysis. However, as social data could contain more information than a formal text, we The output of this step will represent the intermediate sentiment words of each: interpose SW(Yi) and inter- neg SW(Yi)

Step 3 The purpose of this step can be defined as to refine the annotated dictionary: positive posSW(), negative negSW()and neutral neutSW() dictionaries for each Yi . The task of classification of neutral hashtags is difficult and that could affect the result, so that's why we ignored them during the collect. In fact, a tweet

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

that contains a neutral hashtag such as #modi could be either negative or positive. Therefore, we have to construct neutral basing on the word occurrence Occ(wj) for all in the different classes. This will allows us to construct the final dictionaries by using of Algorithm 1.

We conducted empirical test that consists of testing a number of values (between 0.5 and 0.8), in order to constitute the limit that allows classifying sentiment words with the smallest error rate. In our case 0.7 was the best value. At the end, we will assign a score to sentiment words: $1, 0, -1$ for positive, neutral and negative, respectively.

Algorithm 1 Extract Twitter sentiment

```
procedure TWITTER-CONNECTION()
    consumer − key =' xxxxxxxx'
    consumer − secret =' xxxxxxx'
    access − token =' xxxxxxx'
    access − token − secret =' xxxxxxxxx'
    self.auth = OAuthHandler(consumer − key, consumer − secret)
    self.auth.set − access − token(access − token, access − token − secret)
    self.api = tweepy.API(self.auth)
end procedure

procedure TWEET-CLEANING(t)
    tweet = t.remove − Stop − words
    Return tweet
end procedure

procedure TWEET-CLASSIFICATION(t)
    t = Tweet − Cleaning(t)
    tweet − polarity = t.sentiment.polarity
    tweet − polarity
end procedure

procedure GET-TWEETS(q, count)
    fetched − tweets = self.api.search(q = query, count = count)
    Return fetched − tweets
end procedure

procedure MAIN()
    st = SentimentalTwitter()
    tweets = st.fetch − tweets(query =' politics', count = 300)
    PositiveTweets = tweetsthatsentiment =' positive'
    NegativeTweets = tweetsthatsentiment =' negative'

    for tweet t in PositiveTweets do
        print(t)
    end for
    for tweet t in NegativeTweets do
        print(t)
    end for
end procedure
```

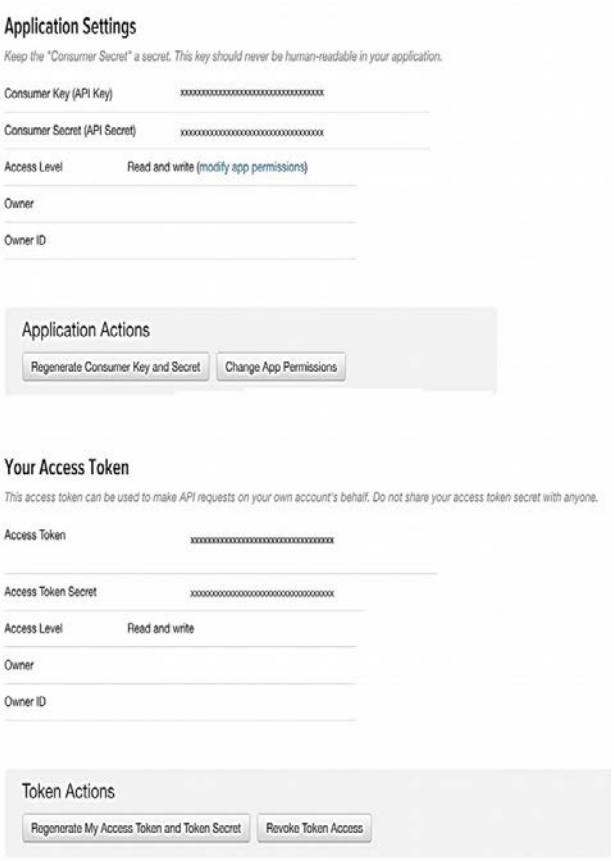


Figure 2 shows the modules of the above shown first stage:

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

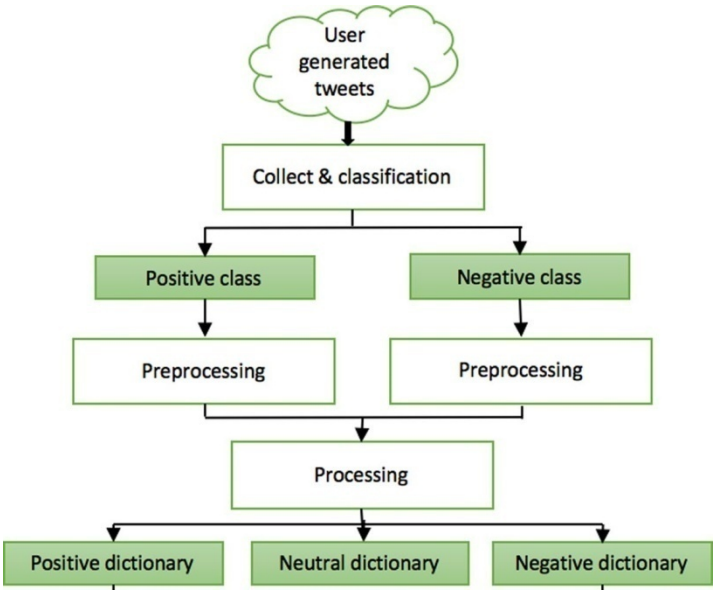


Fig.2 First stage of the model

SECOND STAGE: CLASSIFICATION

In this stage, we will classify new tweets based on the SW dictionary build in the previous stage.

Step 1 Includes the process of collecting and storing new tweets for each Yi separately. Collected data goes through the following number of steps.

Step 2 Includes the process of preprocessing the data as following:

- Removing duplicated data can be defined as a process of removing duplicated tweets in order to avoid misleading results. As a post could include multiple hashtag and it could be extracted for multiple times.
- Tokenization is the same process that we used in the first stage is applied to the new used tweets.
- Handling negation, Here negation words can be defined as words like (no, not, nothing and nonce). These words can significantly affect the overall polarity of a sentence, it is considered as a very important criterion in the field of sentiment classification. As stated in researches we reverse the sentiment polarity of the words that come after a negation word until reaching a punctuation mark.
- Handling repetition is a process in which words are going to be detected that are written in uppercase or constitute more than two of the same consecutive letters.
- Applying morphology by following the same rules as the first stage.

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

Step 3 In this step, we are going to calculate degree of polarity of the tweets based on their semantic orientation of words which are assigned in stage 1. For this, we are going to apply the following two actions:

- Balancing it is worth noting that the used language in social media posts is not conventional and could contain some special words such as those that are written in upper case or contain the repetition of more than two consecutive letter which we call "extended word".
- Calculating polarity degree, in this we going to find the polarity of a tweet which is calculated by adding up the independent score values of words stated in t. m is the length of t.

$$p(t) = \sum_{k=1}^{m} score(w)$$

Step 4 in order to classify tweets, we use p(t) as following: We classify scored tweets into seven classes according to the polarity degree into C+3, C+2, C+1, C0, C−1, C−2, C−3and these classes will symbolize tweets as highly positive, moderately positive, lightly positive, neutral, lightly negative, moderately negative, highly negative classes, respectively). For this, we conduct empirical test to determine the limit of each class. If case1: $0 < p(t) \leq 3$, then the tweet is classified as lightly positive. If case2: $4 \leq p(t) \leq 6$, it is moderately positive. If case3: $p(t) \geq 7$, then, it is highly positive. If case4: $-3 \leq p(t) < 0$, then, the tweet is classified as lightly negative. If case5: $-6 \leq p(t) \leq -4$, it is moderately negative. If case6: $p(t) \leq -7$, then, it is highly negative. There is also a possibility for a sentiment score to be equal to 0, if it is equal to zero, then, the tweet is classified as neutral.

THIRD STAGE: PREDICTION

Several researchers have considered three classes' i.e. positive, negative and neutral classes to determine the sentiment of a document based on the words and/or emoticons and only few ones, such Khatua et.al.

Have examined the polarity degree (i.e. highly, moderately, weakly positive and negative classes). But authors have considered only two indicators i.e. strongly positive and strongly negative.

CLASSIFICATION ACCURACY EVALUATION

To test the capacity of autonomously creating dynamic dictionaries to categorise tweets, we randomly picked 210 tweets from political Twitter corpora: 30 for each

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

class. For each candidate, the tweets were carefully evaluated and categorised as positive, moderately positive, very positive, mildly negative, moderately negative, strongly negative, or neutral. The same data was then processed, as previously stated, though several procedures such as "removing stop words, applying tokenization, stemming, and numerous filters." The preceding stage was completed with the assistance of TreeTagger, a tool for annotating text with part-of-speech and lemma information. TreeTagger has also been enhanced to handle negation, URLs, users, Twitter mentions and hashtags, and intensifiers.

CONCLUSION AND DISCUSSION

Sentiment analysis has been shown to be useful in anticipating people's reactions or opinions by evaluating large amounts of social data on a certain issue. The technique we proposed consists of several steps, beginning with the creation of a dictionary of word polarity based on a very small set of positive and negative hashtags related to a specific given subject, followed by the classification of posts into several classes and the balancing of sentiment weight using new metrics such as uppercase words and the repetition of more than two consecutive letters in a word.

However, the proposed technique still has certain drawbacks. For starters, it is incapable of comprehending emoticons. Second, we exclusively used data from Twitter. Third, we do not have access to big amounts of data for this algorithm. To address these three constraints, we propose a more efficient and global model that can operate with bigger amounts of data.

REFERENCES

3. Hansen LK, Arvidsson A, Nielsen FA, Colleoni E, Etter M. Good friends, bad news-affect and virality in twitter. In: Future information technology, communications in computer and information science. Berlin: Springer; 2011. p. 34–43. https://doi.org/10.1007/978-3-642-22309-9_5.
4. Wang H, Castanon JA. Sentiment expression via emoticons on social media. In: 2015 IEEE international conference on Big Data (Big Data). 2015. p. 2404–8.
5. Mahmood T, Iqbal T, Amin F, Lohanna W, Mustafa A. Mining Twitter big data to predict 2013 Pakistan election winner. In: INMIC. 2013. p. 49–54.
6. Razzaq MA, Qamar AM, Bilal HSM. Prediction and analysis of Pakistan election 2013 based on sentiment

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved

analysis. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014). 2014. p. 700–3.

7. Livne A, Simmons M, Adar E, Adamic L. The party is over here: structure and content in the 2010 election. In: Fifth International AAAI conference on weblogs and social media. 2011.
8. Wicaksono AJ, Suyoto P. A proposed method for predicting US presidential election by analyzing sentiment in social media. In: 2016 2nd international conference on science in information technology (ICSITech). 2016. p. 276–80.
9. Smailovi J, Kranjc J, Grčar M, Žnidaršič M, Mozetič I. Monitoring the Twitter sentiment during the Bulgarian elections. In: 2015 IEEE international conference on data science and advanced analytics (DSAA). 2015. p. 1–10.
10. Stavrianou A, Brun C, Silander T, Roux C. NLP-based feature extraction for automated tweet classification. In: Proceedings of the 1st international conference on interactions between data mining and natural language processing, vol. 1202, DMNLP'14. Aachen: CEUR-WS.org; 2011. p. 145–146.
11. Ramteke J, Shah S, Godhia D, Shaikh A. Election result prediction using Twitter sentiment analysis. In: 2016 international conference on inventive computation technologies (ICICT), vol. 1. 2016. p. 1–5.
12. Tumasjan A. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social media. 2010.
13. Ruths D, Pfeffer J. Social media for large studies of behavior. Science. 2014;346(6213):1063–4.
14. Gayo-Avello D. No, you cannot predict elections with Twitter. IEEE Internet Comput. 2012;16(6):91–4.
15. Balasubramanyan R, Routledge BR, Smith NA. From tweets to polls: linking text sentiment to public opinion time series. Icwsm. 2010; 11:1–2.
16. Chesley P, Vincent B, Xu L, Srihari RK. Using verbs and adjectives to automatically classify blog sentiment. In: AAAI symposium on computational approaches to analyzing weblogs (AAAI-CAAW). 2006. p. 27–9.

Page 99-110 © MANTECH PUBLICATIONS 2022. All Rights Reserved