Distinguishing Reviews Through Sentiment Analysis Using Machine Learning Techniques

S Biruntha [1]
Assistant Professor/
CSE Sri Krishna College of Engineering and Technology,
Coimbatore, India
banupriya12317@gmail.com

Arul Ganeshan S [2]
UG Scholar / CSE
Sri Krishna College of Engineering and Technology,
Coimbatore, India
18eucs012@skcet.ac.in

Ashwin B [3]
UG Scholar / CSE
Sri Krishna College of Engineering and Technology,
Coimbatore, India
18eucs018@skcet.ac.in

Padmasankar K S [4]
UG Scholar / CSE
Sri Krishna College of Engineering and Technology,
Coimbatore, India
19eucs506@skcet.ac.in

2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0816-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACCS54159.2022.9785354

***Abstract* - Opinion Mining is a technique of automated extraction of information from the opinion of people about any certain subject or problem. The feasibility of Opinion mining and Sentiment Analysis tool is to "process a bunch of indexed lists for a given thing, creating a rundown of item credits (quality, highlights and so forth) and accumulating opinion". Yet, with the progression of time additional fascinating applications and improvements appeared around here and presently its fundamental objective is to make PC ready to perceive and produce feelings like human. This paper will attempt to zero in on the essential meanings of Opinion Mining, investigation and Sentiment Analysis is the name given to this new field of study. In recent years, scientists have come up with a number of solutions to the problem. Information Retrieval (IR) and Natural Language Processing (NLP) interact in the subject of Opinion Mining and Sentiment Analysis, which has a few distinct trains, such as message mining and Information Extraction. Sentiment Analysis may be performed using a variety of NLP tools, all of which this paper aimed to demonstrate, from basic definitions to a broad range of applications. Recently, it's been a really active exploring zone. In fact, it has made its way into everything from board science to software engineering of etymological assets needed for Opinion Mining, scarcely any AI methods based on their utilization and significance for the examination, assessment of Sentiment arrangements and its different applications. They are attempting to get assessment data to investigate and sum up the suppositions communicated naturally with PCs. Opinion Mining**

Keywords: Sentiment analysis, Opinion mining, Fake reviews, Language Processing, Information extraction, Textual suppositions.

I. INTRODUCTION

Since the advent of the Internet, people's methods of expressing their thoughts and emotions have changed dramatically. Currently, much of this is accomplished via the use of internet mediums such as blogs, forums, item audit websites, and other social media. Many people these days use social networking sites like Facebook, Twitter, Google Plus, and others to express their views, evaluate, and share insights on their daily routines. This intuitive media is made possible by the use of web-based networks, which allow consumers to communicate with and influence others via gatherings. We're seeing an explosion in the amount of opinion-rich content being generated through web-based media, such as via things like tweets and blog posts, as well as comments and audits. In addition, web-based media allows enterprises the opportunity to connect with their customers in order to publicise their products and services. For the most part, people rely heavily on user-generated material on the internet for independent guidance. For example, if someone has to buy anything or needs help, they'll first look at online reviews and other information about it through web-based media before making a decision. The volume of material generated by clients is too much forthe average customer to keep up with. Because of this, a variety of opinion-gathering methods are in use across the board.

Opinion investigation (SA) informs customers whether or not the information they're about to get about the item is to their taste. Advertisers and businesses use this information to learn about their products or services so that they may display them in a way that meets the needs of their customers. For the most part, solutions for recovering printed data revolve around the actual data that's been lost. Some printed material expresses abstract aspects that are genuine to life. Sentiment analysis's nucleus is made up of a variety of subjective components, such as assessments, views, evaluations, and even sentiments (SA). As a result of the massive growth of web-based data sources like online journals and informal groups, they provide a wide range of possibilities for new applications. Considerations like positive or negative conclusions about the items provided by a suggestion framework, for example, might be predicted by using SA in relation to those things.

II. SENTIMENT ANALYSIS

One way to describe sentiment analysis is as a process

978-1-6654-0816-5/22/$31.00 ©2022 IEEE

that uses Natural Language Processing (NLP) to extract information about people's thoughts and emotions from a variety of sources, such as tweets (NLP). Textual suppositions may be categorised as "positive," "negative," or "unbiased" as part of a sentiment analysis. Subjectivity research, assessment mining, and evaluation extraction are all terms that have been used to describe this process.

III. LITERATURE REVIEW

Many researchers have recently focused their efforts on "sentiment analysis," which is the study of human emotions. A paired structure was originally intended, which would assign audits or findings to bipolar classifications, for example, positive or negative, according to how they were.

According to Pak and Paroubek [1], tweets may be characterised as both good and negative if they are evenly distributed. Using the Twitter API, they gathered tweets and then used emojis to further explain those messages to create a Twitter corpus. They developed an opinion classifier based on the multinomial Naive Bayes algorithm that incorporates features like N-grams and POS-labels into the system. Because it only included tweets with emoticons, the preparation set they used was less successful .

Two models were used to organise tweets by Parikh and Movassate [2], the Naive Bayes bigram model and the Maximum Entropy model. Using the Maximum Entropy model, they found that the Naive Bayes classifiers performed considerably better.

Using far-off management, Go and L. Huang [3] offered a solution for feeling examination of Twitter information based on tweets with emojis that filled in as noisy names. They use Naive Bayes, Max Ent, and Support Vector Machines to build models (SVM). Unigrams, bigrams, and POS make up their component space. Unigrams were thought to be more effective as elements since SVM outperformed other models.

For tweets, Barbosa and coworkers devised a two-stage programed feeling assessment technique. In the second step, the emotional tweets were categorised as either good or negative, based on their content. Retweets, hashtags, connections, accentuation, and interjection markings connected to highlights such as the earlier extremity of words and the POS were all used in the component space.

It was Bifet and Frank [5] that used the Firehouse API to get Twitter streaming data, which provided all messages sent and received by any client and made them publicly available. There were tests for the gullible bayes, the Hoeffding tree, and the stochastic inclination drop. That SGD-based model, when used with a proper learning rate, outperformed the rest of them.

Positive, negative, and nonpartisan opinions may be divided into three categories, according to Agarwal [6] et al. Different models were examined, including unigram models, element models, and tree-part models, among others. In the tree-based approach, tweets were treated as individual branches. More than 10,000 highlights are used in the unigram model, which includes 100 parts. They appear at the conclusion of the highlights that combine the earlier extremities of words with their parts of speech (pos) labels and play a key role in the work of order. The model based on tree parts won over the other two.

To combat overuse, Davidov et al. [7, 8] presented in tweets, the Twitter client used a variety of elements, including accentuation and single words, n-grams, and examples, to describe the opinion type. They used the K-Nearest Neighbor technique to assign marks to each model in the preparation and test set by constructing an element vector.

Using the Twitter API, Po-Wei Liang [8] and colleagues were able to collect data from Twitter accounts. There are three unique classes for their preparation information (camera, film, and portable). Certain, negative, and non-assessments are all terms used to describe this data. The tweets containing conclusions were combed through and analyzed. A unified Naive Bayes model was used, and Naive Bayes improved on freedom presumption. They also used the Mutual Information and Chi Square element extraction methods to get rid of unnecessary components. Finally, it is possible to predict the general direction of a tweet. Positive or bad, for example,

Naive Bayes classifiers by Pablo [9] et al were used to detect tweets that were extreme in English. Baseline and Binary were built to categorise tweets as certain, negative, or neutral, respectively, using a variety of different terminology. Nonpartisan tweets will not be taken into consideration. Classifiers took into account lemmas (items, actions, modifiers, and qualifiers), clarity lexicons, and multiwords from multiple sources, as well as valence shifters.

To examine feelings, Turney et al. used the bag of words approach, in which the relationships between words were not even considered, and archives were treated as a collection of words. An opinion for the full report is formed by taking into account the sentiments of each individual as well as their ability to conglomerate.

The lexical knowledge base WordNet was used by Kamps [12] et al. to determine the enthusiastic essence of a word in different ways. They created a distance measure based on the semantic extremes of descriptors that were still up in the air.

Sentiment classification was achieved using a group system that was created by combining several capabilities and methodologies. In their research, both part-of-discourse data and word relations were used, as were three different types of classifiers (Naive Bayes, Maximum Entropy, and SVM) in their research. To better characterise feelings, they used several methods, such as

fixed mix, weighted blend, and meta-classifier mix.

Provokes and an efficient approach for mining opinions from Twitter tweets were presented by Luo [14] et al. Spam and wildly inconsistent linguistics complicate evaluation recovery in Twitter testing.

IV. EXISTING SYSTEM

Under the current system, there are a number of approaches to analysing a dataset of product reviews. We also demonstrated methods for classifying sentiment in product reviews using two separate datasets using supervised learning. All sentiment categorization algorithms were investigated in our exploratory approaches to find which algorithm was most accurate. Detection methods were unable to discriminate between fraudulent good ratings and bogus negative ones. The following are studies conducted using the current system's document-based opinion mining. The most significant work was completed using "poor" and "incredible" seed words to determine the semantic direction and a point-insight common data technique to calculate the semantic direction. Every phrase in an archive has to be analysed to find the regular semantic direction of emotion. The semantic orientation of context-independent views is established, while the semantic orientation of context-dependent opinions is inferred using linguistic norms. The context-dependent opinions were determined by extracting local data from other reviews that commented on the same product feature.

V. PROPOSED SYSTEM

We might use additional datasets, such as those from Amazon or eBay, and a variety of component selection approaches in the framework we've described. As an additional option, we may use sentiment classification algorithms to identify reviews that are fabricated. Some of these tools may be used to assess our performance in NLP. This system employs an unsupervised, dictionary-based approach. Opinion words and their synonyms and antonyms are determined using WorldNet as a dictionary. To a large extent, this project is an extension of the previously mentioned Mining and Summarizing Customer Reviews project. The "Document Based Sentiment Orientation System" has been described in detail. The approach took into account feedback from customers and reviewers on the items. Using this approach, the total number of papers classified as "positive," "negative," or "neutral" in each archive is shown separately in the final result. As a consequence of the framework's assistance for customers, they can clearly identify the good and bad archives. The majority of opinion words are used to assess the extremes of the presented papers.

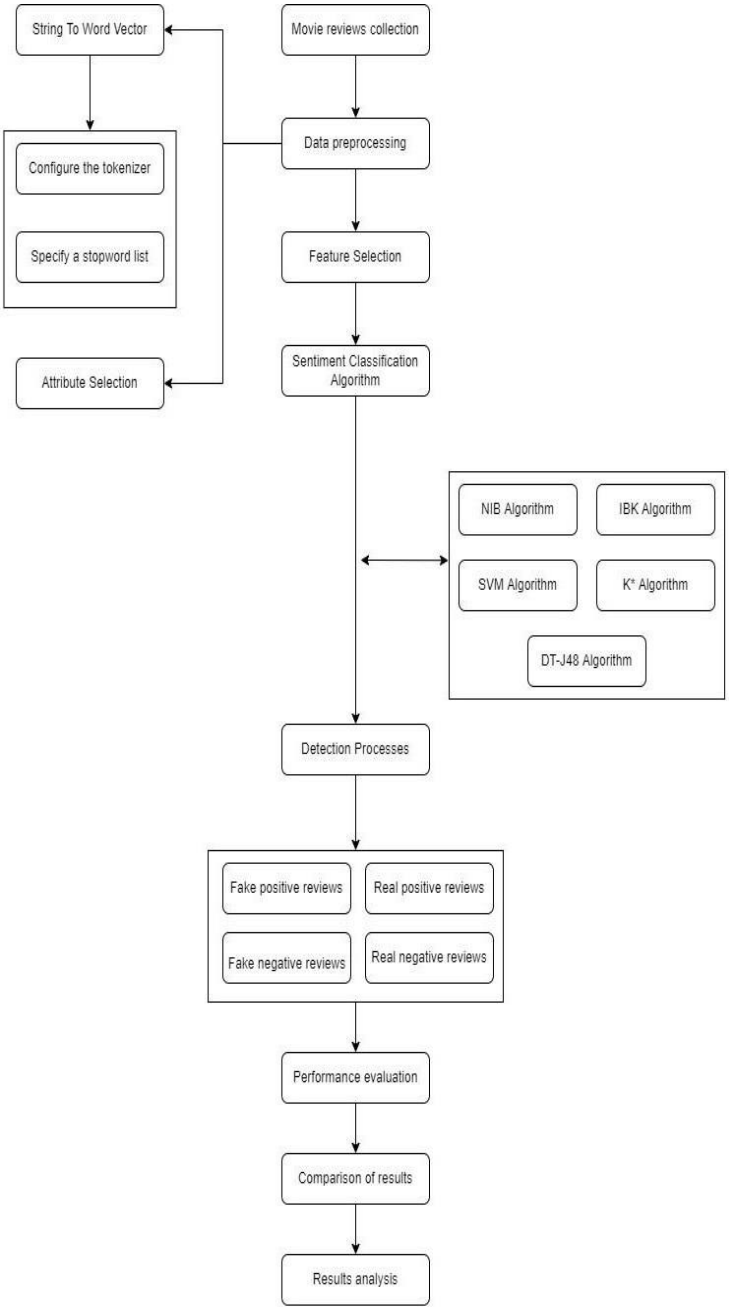

Fig 1 Architecture: Illustrates the proposed system of this research.

VI . MODULES

Data Preprocessing:
URL and Hash tags:

Because the material can only be shared in 140 characters, the user uses URLs and hashtags to link to other relevant resources. Tweets containing this kind of information must be dealt with in a professional manner. In the suggested approach, URLs and hashtags are omitted entirely from tweets.

The mix of capital and lowercase letters in a user's tweet may offer words additional significance if they are arranged in an irregular pattern. As a last step toward removing any potential misunderstanding, the proposed system would lowercase all tweets.

Punctuation and blank spaces are identified and omitted in order to avoid repetition of highlights and distinct challenges. Punctuation and blank areas are omitted.

Document-, phrase-, or sentence-level mining approaches may all be employed to get to the data you're looking for. The technology used for data extraction might be supervised or unsupervised. NB, ME, and SVM are examples of Naive Bayes (NB) and Maximum Entropy (ME) approaches used in supervised learning. Using the unsupervised technique, the suggested system does not utilise information from prior tweets while analysing new tweets. The following are the stages in the proposed system.

Words to avoid:

Various terms are tagged, and some words indicate that these words present in the tweet have no bearing on the sentiment analysis of the tweeter. Stop words are terms that don't contribute anything to the sense of a phrase and should be avoided in tweets. In order to decrease additional complexity, often used terms like "is," "a," and "the," which are labelled as '_IN," '_DT,' '_CC, '_TO,' and '_VBZ' are removed.

Stemming:

Stemming is the process of removing affixes from words to reveal their root form. It's the same as taking a chainsaw to a tree and chopping it down to the trunk. The words are indexed via a process known as stemming, which is carried out by search engines. As a result, search engines can only store the stems of words, rather than all of the variations of a word. It reduces index size and improves retrieval accuracy by using stemming.

NLP is used to categorize words into different components of speech. This is done by splitting the phrase into its grammatical components and marking each word in the sentence with its corresponding extension.

Tagging Output:

Negative and positive datasets are used to compare the POS tagging output. Otherwise, it's a terrible thing if the positive numbers are higher than the negative ones. It is represented as intermediate if the positive and negative values are the same.

VII. ALGORITHM DESCRIPTION

Implementation of the algorithm is described in the following steps.

Step 1: A URL is provided for data extraction.

Step 2: Tokenize the extracted data text and store individual words in an array.

Step 3: A single stop word is read from the stop word list.

Step 4: Then with the sequential search technique, stop words and target words are compared in an array.

Step 5: When it matches, it will remove the word from the array, and the comparison will be continued until the length of the array is reached.

Step 6: After removing a stop word, a new one is read from the stop word list, and the algorithm repeats step 3 again. The process continues continuously until all the stop words have been compared and removed.

Step 7: In addition to displaying the resultant text without stop words, necessary statistics are displayed, such as the number of stop words removed from the target text, the total count of words in the target text, and the individual stop words found are displayed.

Step 8: After the removal of stop words the result is processed again by eliminating affixes from words in order to get to their root by removing stemming word.

Step 9: In the final process, negative and positive datasets are used to compare the POS tagging output to get the number of positive and negative reviews count.

Step 10: And also, the final result shows whether the review is negative or positive by comparing positive and negative review counts.

Performance evaluation of algorithm

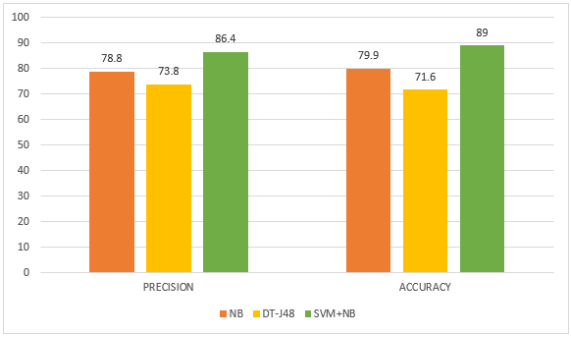

Fig 2 **Precision and accuracy percentage of algorithms**

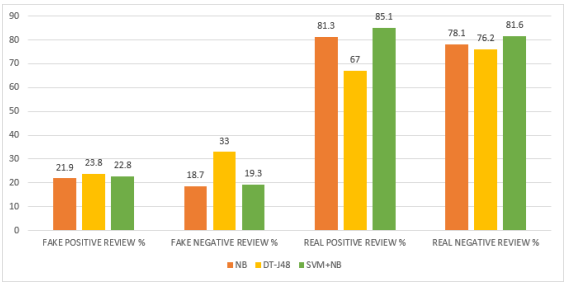

Fig 3 Fake and real percentage of positive and negative reviews

Figure 2 and 3 illustrate that the proposed model algorithm (SVM+NB) is compared with some of the previously used algorithm methods to prove that the precision, accuracy, and positive review percentage of our algorithm are high when compared to others.

VIII. RESULTS

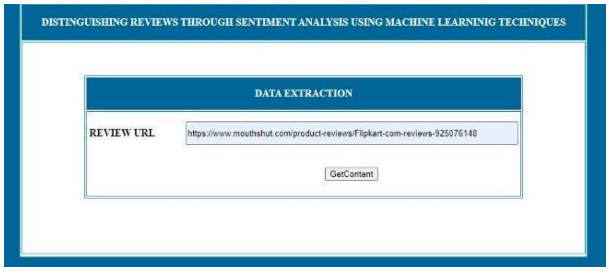

Fig 4 Index page

In Figure 4, the URL should be provided for data extraction.

Fig 5 Stop words

In Figure 5, text was fed into an algorithm and the results obtained from the system are listed. Many stop words were removed, thus reducing the number of words.

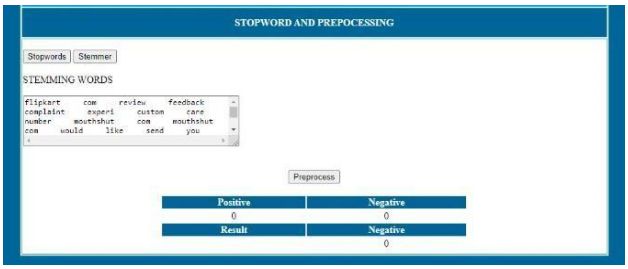

Fig 6 Stemming words

In Figure 6, the results of stop words were fed into the algorithm and the results obtained from the system were listed, by removing a greater number of stem words.

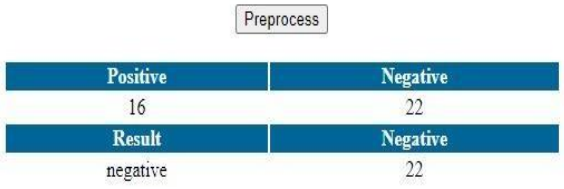

Fig 7 Final Output

Figure 7 shows the positive, negative, and the results obtained.

In the final process, negative and positive datasets are used to compare the POS tagging output to get the number of positive and negative reviews count. And also, the result shows whether the review is negative or positive.

IX. CONCLUSION AND FUTURE WORK

Due to the rapid growth of the internet, the number of reviews of the products/merchandise will rise. Because the Internet generates such vast volumes of data, it's impossible to evaluate the quality of the customer evaluations. Pretend reviews may be written by anybody, and some companies are even paying people to publish them. As a result of the large number of fake internet reviews, the ability to distinguish between genuine and fake is essential. In this study, we've discussed many methods for detecting fake reviews, including unattended, supervised, and semi-supervised methods. Different alternatives, such as linguistic, behavioral, and relative possibilities, have been thoroughly examined throughout this work. In addition, we've evaluated a variety of methods for spotting fake reviews. In addition, we'd like to point out the significant challenges of pretend review detection.

REFERENCES

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, “Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. “Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper,2009.

[4] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in social media, 2011, pp. 30-38.

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volume pages 241- 249, Beijing, August 2010.

[8] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data”, IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1- 494673-6068-5.

[9] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (Sem Eval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171-175.

[10] Neethu M.S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013, at Tiruchengode, India. IEEE – 31661.

[11] P. D. Turney, “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, 2002.

[12] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[13] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[14] Zhunchen Luo, Miles Osborne, Ting Wang, An effective approach to tweets opinion retrieval",Springer Journal on World Wide Web Dec 2013, DOI: 10.1007/s11280-013-0268-7.

[15] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACM International conference on Conference on information & knowledge management (pp. 2079- 2088). ACM,2013.

[16] Pan S J, Ni X, Sun J T, et al. “Cross-domain sentiment classification via spectral feature alignment”. Proceedings of the 19th international conference on World wide web. ACM, 2010: 751-760.

[17] Wan, X. “A Comparative Study of Cross- Lingual Sentiment Classification”. In Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 24-31) IEEE Computer Society.2012

[18] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment Treebank." Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013.

[19] Meng, Xinfan, et al. "Cross-lingual mixture model for sentiment classification." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Volume 1,2012

[20] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M.“Lexicon based methods for sentiment analysis”.