iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48240

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: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Sentiment Analysis and Opinion Mining Applied toScientific Paper Reviews

Tanishka Balkrishna Tiwatane[1], Vrushali Balaji Zampalkar[2], Gauri Sunil Bhavsar[3], Sonali Rajendra Bachute[4]
[1]*Student, Computer Engineering Trinity Acadaemy of EngineeringPune, India*
[2]*Trinity Academy of EngineeringPune, India*
[3]*Professor, Computer EngineeringTrinity Academy of Engineering Pune, India*
[4]*Trinity Academy of EngineeringPune, India*

Abstract: Sentiment analysis and opinion mining is an area that has experienced considerable growth over the last decade.Todo this, natural language techniques and machine learning algorithms are used. This research attempts to determine the feelings, opinions, emotions, among other things, of people on something. A hybrid approach that combines an unsupervised machine learning algorithm with techniques from natural lan- guage processing is proposed to analyze reviews. The first aim of this analysis is to automatically determine the orientation of a review.This article discusses the problem of extracting sentiment and opinions from a collection of reviews on scientific articles conducted under an international conference on computing in northern Chile.

Index Terms: sentiment analysis, hybrid methods

I. INTRODUCTION

Opinions are central to almost all human activities because they are a key influence on people's behavior. Each time a decision needs to be made, humans look for others' opin-ions. In the real world, enterprises and organizations seekto know public opinion about their products and services.In turn, customers want to know others' opinion about a certain product before buying it. In the past, people lookedfor opinions from their friends and family, while organizationsmade polls or organized focus groups. Nevertheless, with the sudden growth of social networks such as Twitter and Facebook, individuals and organizations use data providedby these means to support their decision-making process. The field of sentiment analysis, also called opinion mining, emerged in this context.

There are different techniques for extracting, processing, and seeking objective data in texts. There are different techniques for extracting, processing, and seekingobjective data in texts. These components including opinions, sentiments, and emotions, among others, are the focus of sentiment analysis. Sentiment analysis is an area with great development opportunities, particularly due to the huge growthof data available in the web.

One of the applications of opinion mining is product or service assessment by analyzing user's opinions or reviews. This application is highly important for organizations because it allows discovering what people think and say about a certain trademark. Sentiment analysis includes a great amount of tasks such as sentiment extraction and classification, subjectivity detection, opinion summary,and opinion spam detection. The domain of scientific paper reviews presents some major challenges, such as: 1. Usually classes are unbalanced, because there is a strong bias towards negative opinions. 2. Different reviews usually vary in termsof the number of assessments. 3. Normally, there is not aclear correlation between the number of positive and negative opinions with the final evaluation made by reviewers.

II. LITERATURE SURVEY

1) M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri. In this paper, we propose an integrated social media content analysis platform thatleverages three levels of features, i.e., user-generated content, social graph connections, and user profile ac- tivities, to analyze and detect anomalous behaviors that deviate significantly from the norm in large-scale social networks. Several types of analyses have been conductedfor a better understanding of the different user behaviors in the detection of highly adaptive malicious users. We also collected a significant number of user profiles from Twitter and YouTube, along with around 13 million channel activities. Extensive evaluations were conducted on real-world datasets of user activities for both social networks. The evaluation results show the effectiveness and utility of the proposed approach.

©IJRASET: All Rights are Reserved

2) I.-R. Glavan, A. Mirica, and B. Firtescu.[4] This study conducts an analysis on how social media is used by Offi cial Statistical Institutes to interact with citizens and disseminate information. A linear regression technique is performed to examine which social media platforms (Twitter or Facebook) is a more effective tool in thecommunication process in the offi cial statistics area. Our study suggests that Twitter is a more powerful tool than Facebook in enhancing the relationship between offi cial statistics and citizens, complying with several other studies. Next, we performed an analysis on Twitter network characteristics discussing "offi cial statistics"using NodeXL that revealed the unexploited potential of this network by official statistical agencies.

3) H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng, and T. S. Chua. In this paper, we find that users stress state is closelyrelated to that of his/her friends in social media, and we employ a large-scale dataset from real-world social platforms to systematicallystudy the correlation of users' stress states and social interactions. We first define a set of stress-related textual, visual, and socialattributes from various aspects, and then propose a novel hybrid model - a factor graph model combined with Convolutional NeuralNetwork to leverage tweet content and social interaction information for stress detection. Experimental results show that the proposedmodel can improve the detection performance by 6-9percent in F1- score. By further analyzing the social interaction data, we also discoverseveral intriguing phenomena, i.e. the num- ber of social structures of sparse connections (i.e. with no delta connections) of stressedusers is around 14percent higher than that of non-stressed users, indicating that the social structure of stressed users' friends tend to beless connected and less complicated than that of non-stressed users.

4) R. L. Rosa, D. Z. Rodr ıguez, and G. Bressan.[15] From this section Online social networks like twitter have a ton of data. However, regularly individuals don't give individual data, like age,sexual orientation and othersegment information, albeit the certainty investigationuses such data to foster valuable applications in individ- uals' day to day routines. This exploration shows that quite possibly the most significant parameter containedin the client profile is the age bunch, which shows that there is ordinary conduct among clients of as old as, particularly when these clients expound on. With a similar theme Detailed examination with 7000 sentences has been directed to figure out which elements are significant, like the utilization of accentuation, number of characters, sharing of media, different subjects, and which ones can disregard the age bunch grouping. Diverse learning ma- chine calculations have been tried for the characterization of juvenile and grown-up gatherings and the Word2Vec has the best exhibition with exactness up to 0.95 in the approval test.

5) R.B. Liu.[16] In this research system are trying to fillthe gap between emotional recognition and emotional co-relation mining through social media reviews of natural language text. The association between emotions, rep- resented as the emotional uncertainty and evolution, is mainly triggered by cognitive bias in the human emotion.Three different types of features and two deep neural- network models are provided to mine the emotion co- relation from emotion detection using text. The rule on conflict of emotions is derived on a symmetric basis. TF- IDF, NLP Features and Co-relation features has used for feature extraction as well as section and Hybrid deep learning algorithm for classification has used to demon- strates the entire research experiments. Finally system evaluates the performance with various existing system and show the effectiveness of proposed system.

III. METHODOLOGY

Opinion mining are related to data extraction and pre- processing, natural language processing, and machine learning methods, which play a vital and crucial role in the task of de- termining the aspect of an of an opinion. A learning task may be divided into two broad categories: Supervised learning, in which classes are provided in data, and unsupervised learning, in which classes are unknown and the learning algorithm needs to automatically generate class values. Supervised methods such as na¨ıve Bayes and Support Vector Machines were used. For the unsupervised learning task, an idea based on part-of- speech tagging and keyword matching was used. Furthermore,a hybrid technique which combines both supervised and un- supervised methods is proposed. Deep learning methods have not been tested cause of the small size of the dataset. Since deep learning methods perform well in sentiment analysis the number of arguments that must be estimated for deep learning to work fine is too big for the amount of data present inthis dataset. Enlarging the data set is a difficult task because scientific reviews are an occluded genre and as such getting access to more data is not easy. Collecting more reviews has been left for future work, and the application of deep learning methods on this dataset has been left for future work. Paper reviews are represented in a structured format using json. As part of the preprocessing step the raw data has been checked manually and corrections have been applied where required. After reading the corrected data, another preprocessing step is needed before constructing th supervised and unsupervised classifiers. All the classifiers generate a report in text format that is visualized by the final user.

A. Supervised Learning

As for SVM, this approach has a proved theoretical basis and has empirically shown to be the most perfect and accurateclassifier for text documents. The classifier developed by Pyhton scikit-learn library, libsvm implementation was used. Specifically, a linear kernel was used cause it gave betterresults than others available in the library. The optimal classi- fier parametrization was obtained via empirical tests. Default parameters were used for the other configurable elements of the implementation because they provided better results. For SVM, an output coding based on error correction codes was used. This method is developed in sklearn libraries andits performance was better than the one other than all the approaches used by default for the implementation, obtaining a 10% improvement in terms of the average metric F1-score. The selected code size is twice bigger than the amount of classes. This parameter was selected via empirical performance eval- uation (values from 0.5 to 3.0, with 0.25 increments were tested). In both cases, the training of the classifiers was done by splitting the data set into a training set and a testing setwith a 70% and 30% proportion, respectively main values, this representation being SVM input. SVD is applied in order to reduce dimensionality, even though SVMis not sensitive to high dimensionalities, this reduction will reduce the computational cost of the method. In the case of POS Tagging neither punctuation marks nor stopwords are eliminated because they contain useful data for the classifier (for example, negation). The text is then entered into Stanford POS Tagger in order to identify its semantic structure. Finally,a manual review is made to look for words (i.e. iterating over each word in the document) found in certain dictionaries so as to mark these instances with additional tags. This list of tokensand their associated tags corresponds to the unsupervised classifier input.

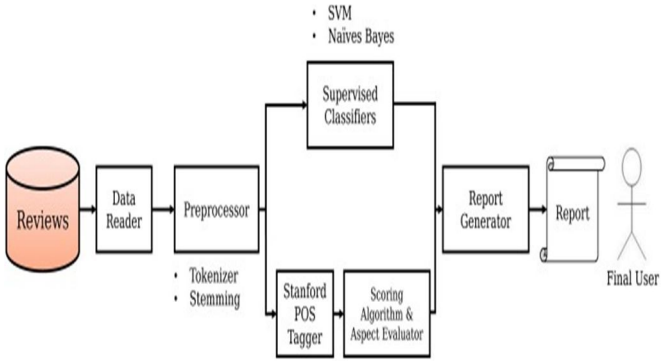


Fig. 1. High level diagrams of the Implemented Methods

B. Unsupervised Learning

Once the text is separated in tokens, the next step is usually made to conduct a morphosyntactic analysis to identifycharacteristics. For example, its grammatical approach. This analysis is known as Part-Of-Speech (POS) tagging. The method uses a text in a given language as an input and throughthe application of its internal POS tagging model, assigns a grammatical approach to the words in a sentence, per say,verb and adjective, among others. In addition, each category has its own features. The complexity of this task is based on the target language to be tested. For example, Spanish is more complex as to verb conjugation and implicit subjects. To applythis technique, pre-processing, stemming is skipped becauseit may prevent obtaining the correct grammar structure. POS tagging proposes two main difficulties: The first one is word ambiguity, which depends on the type of the sentence analyzedand the second one is assigning a grammatical type to a word when the system does not know how to do it. To solve twoof these problems, the context around the word in a sentenceis typically considered and the most probable is selected. The grammatical category has a relevant characteristic. A word belonging to the same word group can replace a token withthe same grammatical category, without affecting the sentencegrammatically.

C. Preprocessing

Before classifying a text, it is necessary to process it. First, punctuation standardization is done, so that writing rules can be respected (for example, “The writing is awful,but the form is correct.” would become “The writing is awful, but the form is correct.” (now, there is a space after the comma)). Once this is done, the text is tokenized, separating it into sentences (according to the use of periods) and each sentence into words. Depending on each case, different preprocessingis done on these words. A TF-IDF scheme is applied to the input text, this representation being Bayes classifier input. A TF-IDF scheme is applied to the input text; then, the Singular Value Decomposition (SVD) method is applied, keeping 100

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

IV. ALGORITHM

A. *Scoring Algorithm*

To evaluate a review, Scoring Algorithm is used over each sentence and then the mean of all the sentences in the review are estimated. The value produced by this scoring Algorithm gives the semantic orientation of the review in terms of a continuous numeric scale. This solution must be discretizedto obtain the classification in the corresponding classes. The binary classification method (classes "1" and "1"), ternaryclassification (classes "1", "0", and "1"), and 5-point scale multiclass classification (from "2" to "2") were tested, obtain- ing different performances in each case cause their increasing complexity. The algorithm was developed by following a rule- based scheme, according to the semantic characteristics of words. Specifically, a dictionary-based approach combined with a series of heuristics was used, these heuristics consist of rules that define the effect of each type of word on the semanticorientation of a sentence. First, each word is tested to be taggedaccording to its semantic characteristics (POS Tagging). In addition, the dictionaries mentioned previously were used to add other tags in each word. The dictionaries are listed below, they were used in order to specify the effect of each wordon the semantic orientation of the sentence. Particularly, the general effect on the sentence, according to a series of pre- established rules, is calculated, depending on the word found and its semantic orientation.

B. *Steps for Score*

1) function SCORESENTENCE
2) TotalScore =
3) PreviousTokens(2) = None
4) Inverted = False
5) TokenScore = 0
6) for all (Token token in TokenList) do
7) Tags = GetTags(Token)
8) TokenScore = GetSentiWordNetScore(Token, Tags)
9) if IsPositive(Tags) then
10) TokenScore = TokenScore * PosBias
11) else if IsNegative(Tags) then
12) TokenScore = TokenScore * NegBias
13) end if
14) if Token == '?' then
15) TokenScore = QMOrientation
16) Next Token
17) end if
18) if IsSuggestion(Tags) then
19) TokenScore = SuggestionOrientation
20) end if
21) if IsInversion(Tags) then
22) Inverted = ¬ Inverted
23) end if
24) if Inverted then
25) TokenScore = TokenScore
26) end if
27) if IsVerb(Tags) and ContainsNo(PreviousTokens) then
28) TotalScore = TotalScore NegatedVerbOrientation
29) end if
30) if IsIncrement(PreviousTokens) then
31) TokenScore = TokenScore * ModFactor
32) end if
33) if IsDecrement(PreviousTokens) then
34) TokenScore = TokenScore/ModFactor

©IJRASET: All Rights are Reserved

35) end if
36) if IsAdversative(Tags) then
37) TotalScore = TotalScore * AdversativeWeight
38) end if
39) TotalScore = TotalScore + TokenScore
40) Update PreviousTokens
41) end forreturn TotalScore
42) end function

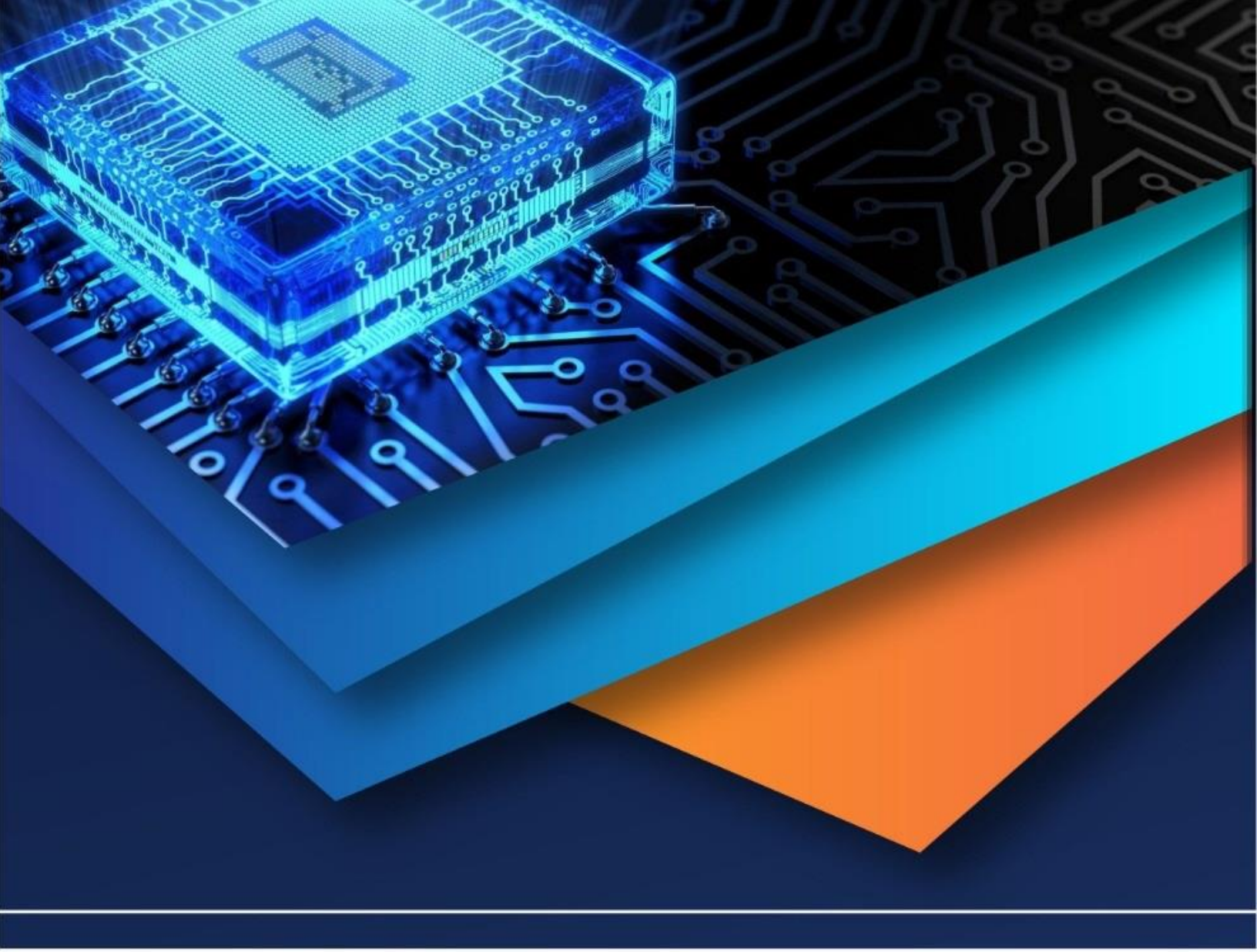
V. CONCLUSION

Concerning the experimental results, it is necessary to enlarge the list of features , so that classifiers perform better and improved classification results are acquired. Also, expanding the data set with more reviews would be very useful in future research, since the current data set is too small to apply some techniques to perform well. This may allow a better evaluation of papers since it would be possible to recognize that a reviewer is strict or not.

Finally, since there are no other papers the proposal in this study is a contribution and innovation for the field of sentimentanalysis .

REFERENCES

[1] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, "Leveraging analysis of user behavior to identify malicious activities in large-scale social networks," IEEE Transactions on IndustrialInformatics, vol. 14, no. 2, pp. 799–813, Feb 2018.
[2] I.-R. Glavan, A. Mirica, and B. Firtescu, "The use of social media for communication." Official Statistics at European Level. Romanian Statistical Review, vol. 4, pp. 37–48, Dec. 2016.
[3] H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng,and T. S. Chua, "Detecting stress based on social interactions in social networks," IEEE Transactions on Knowledge and Data Engineering, vol.29, no. 9, pp. 1820–1833, Sept 2017. .
[4] R. L. Rosa, D. Z. Rodr´ıguez, and G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks," IEEETransactions on Consumer Electronics, vol. 61, no. 3, pp. 359–367, Oct 2015.
[5] R. Rosa, D. Rodr, G. Schwartz, I. de Campos Ribeiro, G. Bressan etal., "Monitoring system for potential users with depression using sen- timent analysis," in 2016 IEEE International Conference on Consumer Electronics (ICCE). Sao Paulo, Brazil: IEEE, Jan 2016, pp. 381–382.
[6] B. Liu, "Many facets of sentiment analysis, a practical guide to sentimentanalysis," Springer International Publishing, pp. 11–39, Jan 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)