iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** X **Month of publication:** October 2017

DOI: http://doi.org/10.22214/ijraset.2017.10130

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:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

Lexicon Based Sentiment Analysis of Twitter Data

Dibakar Ray [1],
[1] *National Informatics Center Kolkata, West Bengal, India*

Abstract*: Sentiment Analysis of twitter data is an active area of Natural Language Processing research. This study explores a unsupervised lexicon based approach to calculate polarity of tweets fro a publicly available twitter corpus. Along with lexicon based search of sentiment bearing words, several rule based methods are used to get the final polarity count of tweets. This study takes into account effect of negation, capitalization, multiple punctuation, slang, and degree modifier
Keywords*: Sentiment analysis; Unsupervised learning; lexicon; Sentiwordnet

I. INTRODUCTION

Proliferation of new technologies, such as mobile devices and social media platforms have made it very easy for people with access to Internet to express their opinion on varied subjects. Participation of millions of people in social media has resulted in a huge amount of data. Such huge data provides a unique opportunity for researchers and business to know about peoples feeling about various subjects around them. Sentiment analysis or Opinion Mining is one such tool to extract sentiment from the social media data. Liu and Zhang [1] define Sentiment Analysis (SA) as the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. Micro blogging website Twitter with its hundreds of millions of users and over 500 million Tweets [2] being sent each day, holds a unique position among social media platforms. Given the sheer volume of the twitter database and being available mostly in public, Twitter data has been extensively used for analyzing people's sentiment about events, products, and services, political views.

Great amount of research is going on about extracting sentiment from social network messages, online reviews, news article [1], [3], [4]. Given its 140 character message length constraints and due to its conversational nature and lack of conventional orthography, Twitter data provides additional challenge for sentiment analysis [5].

Broadly, there are two types of methods for sentiment analysis: supervised and unsupervised learning. Supervised learning approach where sentiment of the object is classified into any of the tree classes- positive, negative and neutral. Supervised learning requires training and testing data. Training data needs to be labelled positive, negative or neutral. Supervised learning methods like naive Bayesian classification, and support vector machines (SVM) are most commonly used by the researchers [1]. Generally, supervised learning methods produces result of high accuracy, but are very domain specific. Accuracy reduces drastically as domain of test data moves away from the domain in which the model was trained. Quality of training data also pay an important role in performance of the supervised learning, high quality labeling of training data often requires huge effort.

Unsupervised learning methods use opinion words and phrases from individual items under study and are than matched against a list of words or lexicon, where each word has a specific sentiment. In this paper we present a Lexicon based sentiment analysis method, and explore how adding different features like capitalization, multiple punctuation, elongation etc affects polarity score.

II. RELATED WORKS

Sentiment analysis is a very active area of NLP research. Sentiment analysis studies are mainly done in the domain of movie and product review [6][7]. Pang, Lee and Vaithyanathan [8] applied various machine learning techniques and bag of features framework for sentiment classification of movie review data. The authors found performance of Naïve Bayes among the worst whereas performance of SVMs found to be the best among the supervised learning methods they have used in their study. Similarly Cui, Mittal & Datar [9] found classifiers such as SVMs more appropriate for sentiment classification, however when training data set size is small, a Naive Bayes classifier might be more appropriate as SVM requires large data set to build a high-quality classifier. In a recent work, Pham and Lee [10] used a Neural Network based model for sentiment analysis of hotel reviews and found that performance of their model is superior to that of probabilistic rating regression method or the frequency-based method. Kanger and Bathla [11] used Neural Network to classify blog data and found very encouraging result. Similarly a number of researchers have worked with Lexicon based methods. Lexicon based methods calculate sentiment polarity as a function of sentiment bearing words in twitter [7], [12], [13]. Khan et al. [14] have used a lexicon backed rule based classification scheme to classify user reviews. The system integrates effect of emoticons, modifiers, negations etc to the lexicon based framework to improve performance. .

Quite a few review of sentiment analysis are published by researchers. Pang and Lee [15] in their review covers techniques and approaches and challenges of Sentiment Analysis. Liu [16] in his book provides an in depth introduction to Sentiment Analysis and presents a comprehensive survey of all important research topics and the latest developments in the field. Medhat et al. [17] have done a meta analysis of 54 published articles to illustrate trend in research of the sentiment analysis applications and algorithms. Hailong et al. [18] in their study used an evaluation metrics for the comparison of existing techniques for opinion mining which includes machine learning and lexicon-based approaches.

Like movie and product review, research on sentiment analysis of Twitter is also gaining ground [19],[20],[21],[22].[23]. Lots of those studies are based on supervised learning methods like Artificial Neural Networks [24] , Distant Supervision method [19]. Most important lexicon based sentiment analysis with twitter corpus is done by Hutto and Gilbert [25]. They combined lexical features with grammatical and syntactical rule based model to classify tweets. The model performed better than the benchmarks. Saif et al.[26] developed a lexicon based model for sentiment analysis of tweeter, which takes into consideration co-occurrence patterns of words in different contexts in tweets and adjust accordingly their pre assigned strength and polarity in sentiment lexicons.".

III.DATA

Data used for this study contains a set of 498 labelled tweets taken from http://help.sentiment140.com/for-students. Data was collected by the owner of dataset using Twitter API and classified based on emoticons [19]. The dataset contains 182 positive, 177 negative and 139 neutral tweets. The data is available at the site in CSV format with emoticons removed. From the csv file which has six fields, field important for this and future studies are extracted. They are – Polarity, Tweet Id, Date of the tweet, User and The Text of the Tweet. Data has three distinct polarity values – negative represented as 0, neutral represented as 2 and positive, represented as 4.

IV.LEXICAL RESOURCE

Unlike supervised learning methods, Lexicon based approaches relies on sentiment words in documents and sentiment orientations of the words are measured by matching these words against content of dictionaries. Such dictionaries contains words along with its associated semantic orientation. Several prior polarity sentiment lexicons are mentioned in research used for sentiment analysis. Most prominent among them are SentiWrdNet, Subjectivity Lexicon, Taboada's Lexicon, MPQA subjectivity Lexicon by Bing Lu etc. For this work SentiWordNet lexicon is sued.

SentiWordNet (SWN)[ii] : SentiWordNet (current version is 3.0) is a publicly available lexical resource designed especially for Sentiment Analysis. SentiWordNet is based on WordNet Each synset is associated to three numerical scores positive, negative and neutral (objective). The sum total of these scores is 1[27] SentiWordNet is found to be most widely used publicly available lexical resource for Sentiment Analysis studies [28].

V. DATA PREPROCESSING

Twitter imposes a restriction on its size, it can only be 140 character long. Hence twitter users' needs to squeeze opinion within those 140 characters with generous use of, slang, emoticons, URLs for reference, and genre specific terminology and abbreviations, etc. [29]. In order to extract meaningful content from tweets , therefore pre-processing of tweets in terms of removal/conversion of such features is required. For this work following pre-processing are done - URL, User tag, Hashtag removal, Tokenization and Part of Speech Tagging (POS) tagging.

A. URL, User and Hash tag Removal

Using regular expression URL, User (with @Prefix) and Hashtag (with # prefix) are removed from the tweet.

B. Tokenization and POS Tagging :

For tokenization we started with Python based nltk.tokenize package. It has a twitter aware module named nltk.tokenize.casual. However we found this tokenizer fails to recognize the multi-punctuation boundary and treats each punctuation of a multi-punctuation occurrence as separate tokens.

Our experiment with Tweet NLP Package from Carnegie Mellon University (CMU)[iii], resulted in tokens as per our expectations. Hence for our experiment CMU, Tweet NLP Package was used for tokenization and POS tagging. The output was stored in a sqlite table for further processing.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

VI. EXPERIMENT

Pre-processed and POS tagged tweeter tokens are then passed through Intensifier detection and Negation Detection processes.

A. Intensifier detection

For this experiment four kinds of intensifiers are considered. Capitalization, Sequence of Repeated Letter, Repeat punctuation and Degree modifier.

B. Capitalize (CP)

Capitalization, of all characters among non capitalized words enhances sentiment polarity of the word. For example For example, "The movie is GREAT!" conveys more intensity than "The movie is great!"[25]. In our experiment when a all capitalize word is detected in a not all capitalize sentence.

C. Multiple Punctuation (MP)

Use of multiple consecutive punctuation marks (e.g. !!!!), conveys sentiment of higher intensity than a word which is not followed by such punctuation marks. For example, "In the new dress she looked She looked fabulous!!! " is more intense than "In the new dress she looked she looked fabulous". In our system, if MP is detected, previous word in our token list is marked as being intensified by MP.

D. Sequence of repeating letters (RL)

Use of sequence of repeating letters, increases magnitude of intensity of the sentiment conveyed through the word. For example "llooovveeee" is less intense than its normal form "love".
Original word for such token is identified by first reducing such repeat letters to maximum two and then searched in a dictionary, if not found such repeat letters are reduced to one and searched in the dictionary. For our study, python library Enchant is used.
In our system CP, MP and RL are handled in similar manner. Polarity of the word being intensified is increased by the polarity of the word "very" as suggested by[30] .

E. Degree Modifiers (DM)

Degree modifiers or booster words or degree adverbs intensifies or decreases the intensity of the sentiment words. Some examples of such intensifiers are 'really', 'very', 'extremely', 'most'. For example 'Food is extremely good at this restaurant' conveys more positive sentiment than 'Food is good at this restaurant'. Similarly 'Service here is bad' is less intense than 'Service here is pretty bad'.
For our study, we have used the intensifier lexicon developed by Taboada et al. [13] and the word which is being affected by the intensifier is identified by searching forward from the degree modifier till punctuation mark is encountered. A percentage scale associated with the degree modifier is then used to calculate the effect [13].

F. Negation (NG)

Negation terms, like 'not', 'never', 'barely' etc. have influence on the overall polarity of the sentiment bearing terms. Research shows that, by properly accounting for the negation increases accuracy of the sentiment analysis [31] .
Determining the scope of negations and how to calculate effect of negation on polarity are subject of intense research [13], [31], [32]. For simplicity we have used Python nltk mark_negation library for negation detection.
The most common approach adopted by researchers to calculate the effect of negation is by flipping of the polarity of the affected sentiment words. Though flipping of sentiment works fine for most of the cases, it fails in certain cases. For example, the term 'Excellent' being one of the most positive sentiment bearing terms, just flipping its polarity when preceded by not (e.g. not excellent) will not be correct. This is because 'not excellent' does not convey most negative sentiment [13].
Hence, for our study, as suggested by Taboada et al. [13] we shift the polarity of the sentiment word toward the opposite polarity by a fixed amount 4. For example, if score for the term terrific is +5 and terrible is -5, then overall score for the sentence - She's not terrific is $5 - 4 = 1$. Similarly, score for the sentence She is not terrible either is $-5 + 4 = -1$,

G. Slang (SL)

Slangs like OMG, LOL are very common in social media interactions. Given the number of character constraints, use of such slangs are very widespread in the twitter world. Hence it is expected that inclusion of such slangs on overall sentiment calculation of the tweets would increases accuracy of the analysis. For this study, list of internet slangs were collected from Webopedia3. Polarity of these slangs were assigned by matching with SlanSD available at KDnuggets[iv] . The dictionary thus greeted contains 553 slangs with f five possible values: - 2 - strongly negative; -1: negative; 0: neutral; 1: positive; 2: strongly positive.

For this work, sentiment bearing tokens with POS tagged Adjective, Adverb, Verb and Nouns are used for Sentiment scoring [12], [33]. Only selected POS tagged terms are used to search SWN.

In order to have parity with other scoring (negation, intensifier, slang etc), SWN scores are converted to a scale of 0 to 5 based on SWN average score according to the following condition –

if swnscore <= 0.20 then swnscore =1
if swnscore > 0.20 and swnscore <= 0.40 then swnscore = 2
if swnscore > 0.40 and swnscore <= 0.60 then swnscore = 3
if swnscore > 0.60 and swnscore <= 0.80 then swnscore = 4
if swnscore > 0.80 then swnscore = 5

If multiple senses are retrieved for a term and for a specific POS then the arithmetic mean of the scores is used to computer sentiment score of the term.

Object score of the term is calculated using the equation

objscore =1-(posscore + negscore) where posscore is positive score, negscore is the negative score of the term. Final term score (termscr) is calculated based on the following condition –

if (posscr> max(negscr, objscr)) then termscr= posscr
if (negscr> max(posscr objscr)) then termscr= negscr
if(max(posscr,negscr,objscr)=objscr)then termscr=0
if(posscr=negscr)then termscr=0

Overall tweet score is calculated by adding the term scores of the tweet.

VII. RESULT

In our study two sets of twitter polarity scores are calculated. For first experiment (EXP1), twitter score are calculated from based on polarity scores based on SWN. No other effects were taken into consideration for EXP1.

For the second experiment (EXP2) along with term scores from SWM, effect of negation (NG), intensifier/diminisher (DM), multi-punctuation mark (MP), capitalization (CP), sequence of repeating letters (RL) etc. are taken into consideration. Different effects are calculated as discussed under the section 'Experiment' and the resultant score is added/subtracted with the tweet score to get the final polarity score.

Performances for the studies are measured by means of precision, recall and F-scores as shown in Table 1.

Table 1. Performance measures for Exp1 and exp2

Performance Measures	Exp1	Exp2
Accuracy	47%	55%
Precision	0.60	0.82
Recall	1.0	0.90
F – Score	0.75	0.85

The result shows that non addition of non lexical parameter improves the performance. High precision in case of experiment 2 show less false positive. Although accuracy for either case is not very impressive, experiment 2 validates finding of [25] that incorporating rules that embody grammatical and syntactical conventions, with lexicon based search outperforms performance with simple lexicon based approach.

Absence of emoticon could be the reason for not so impressive accuracy. As the corpus is generated based on positive and negative emotion search, it can be expected that emoticons are the dominating sentiment expressing features in those tweets and removing them has made it susceptible to incorrect polarity detection .

Other than incorporation of grammatical and syntactical feature, the study shows that incorporating typical twitter feature like slang also increases accuracy. For example, when slang is not incorporated while calculating sentiment score of tweets, accuracy of the

system comes down to about 54%. Incorporating emoticon in the sentiment score calculation is expected to increase the performance as well. However as the twitter corpus is stripped off emoticon that test could not be performed.

VIII. CONCLUSIONS

In this paper a preliminary study on Lexicon based Sentiment Analysis of Twitter data is presented. Corpus used here is a publicly available twitter corpus which are stripped of emoticons. On the twitter data a SentiWordNet based search is performed to find out word level polarity. Word level polarities are then combined with rule based method to calculate the overall polarity.

Despite it being a work in progress, it gives valuable insight into the world of Sentiment Analysis in terms of application development to take care of various rules like negation, capitalization etc. and use of POS tagging library and SenWordNet Lexicon. This work also provide us the framework to extend the study of other data set with more meaningful usage.

REFERENCES

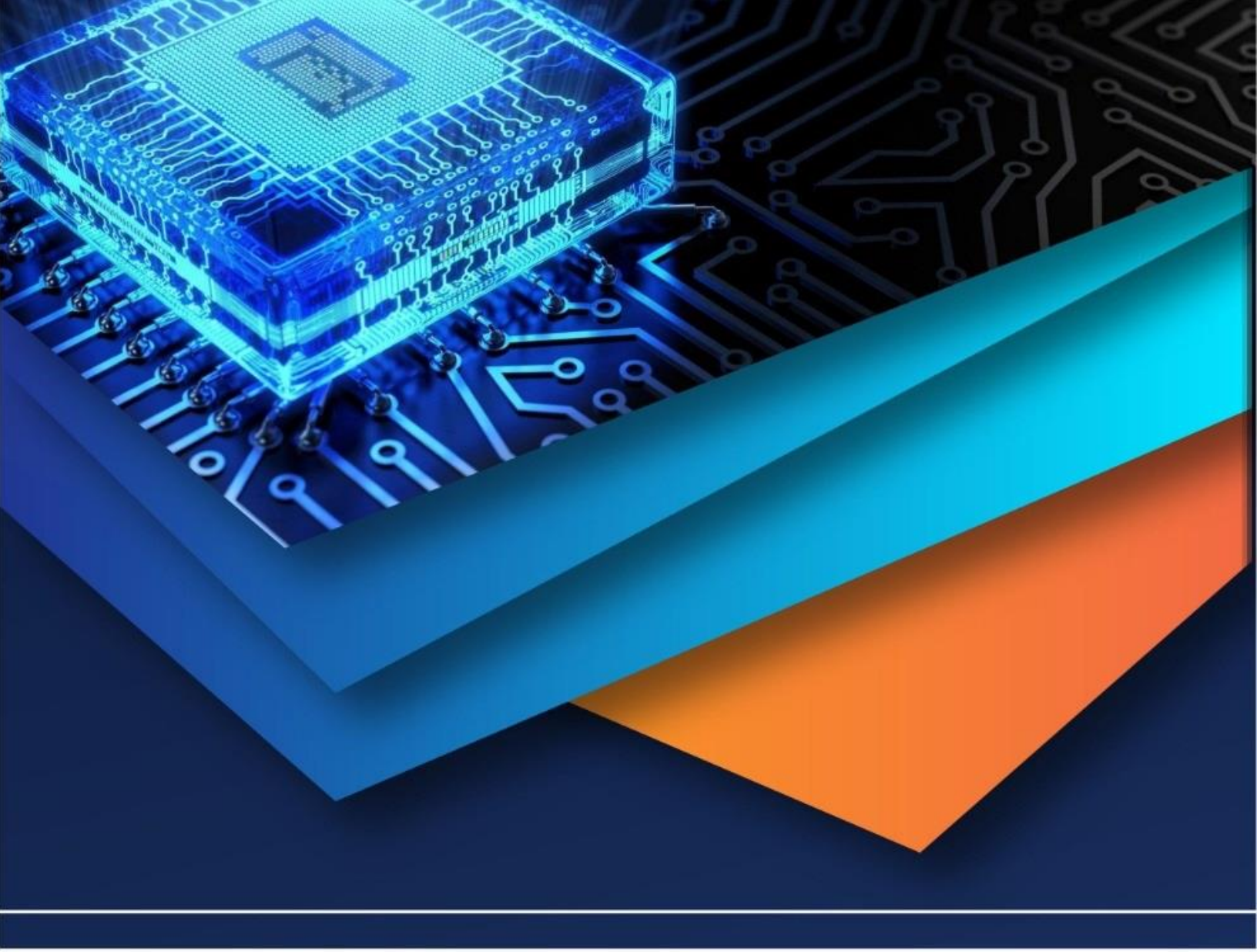
[1] B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis," in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 2012, pp. 415–463.
[2] "Twitter basics," Twitter. Twitter.
[3] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 815–824.
[4] Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 1397–1405.
[5] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," inProceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: shortpapers-Volume 2, 2011, pp. 42–47.
[6] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 519–528.
[7] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, 2002, pp. 79–86.
[9] H. Cui, V. Mittal, and M. Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," in Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, 2006, pp. 1265–1270.
[10] D.-H. Pham and A.-C. Le, "A Neural Network based Model for Determining Overall Aspect Weights in Opinion Mining and Sentiment Analysis," Indian Journal of Science and Technology, vol. 9, no. 18, 2016.
[11] . Kanger and G. Bathla, "Recognizing Emotion in Text using Neural Network and Fuzzy Logic," Indian Journal of Science and Technology, vol. 10, no. 12, 2017.
[12] X. Ding, B. Liu, and P. S. Yu, "A Holistic Lexicon-based Approach to Opinion Mining," in Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 231–240.
[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Computational Linguistics, vol. 37, no. 2, pp. 267–307, 2011.
[14] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PLOS ONE, vol. 12, no. 2, pp. 1–22, 2017.
[15] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.
[16] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.
[17] 2014 11th Web Information System and Application Conference, 2014, pp. 262–265.
[18] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, no. 12, 2009.
[19] A. Bermingham and A. F. Smeaton, "Classifying Sentiment in Microblogs: Is Brevity an Advantage?," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 1833–1836.
[20] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.," in LREC, 2010.
[21] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, AAAI Press, 2011, pp. 538–541.
[22] H. Saif, Y. He, and H. Alani, "Semantic Sentiment Analysis of Twitter," in The Semantic Web – ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 508–524.
[23] D. Zimbra, M. Ghiassi, and S. Lee, "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks," in Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016, pp. 1930–1938.
[24] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." 2014. \
[25] H. Saif, Y. He, M. Fernández, and H. Alani, "Contextual semantics for sentiment analysis of Twitter.," Inf. Process. Manage., vol. 52, no. 1, pp. 5– 19, 2016.
[26] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in LREC, 2010.
[27] A. Das and S. Bandyopadhyay, "SentiWordNet for Indian Languages," in Proceedings of the Eighth Workshop on Asian Language Resouces, 2010, pp. 56– 63.
[28] T. Singh and M. Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis," Procedia Computer Science, vol. 89, pp. 549–554, 2016.

[29] . Muhammad, N. Wiratunga, and R. Lothian, "Context-Aware Sentiment Analysis of Social Media," in Advances in Social Media Analysis, M. M. Gaber, M. Cocea, N. Wiratunga, and A. Goker, Eds. Cham: Springer International Publishing, 2015, pp. 87–104

[30] Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak, "Determining negation scope and strength in sentiment analysis," in 2011 IEEE International Conference on Systems, Man, and Cybernetics, 2011, pp. 2589–2594.

[31] C. Diamantini, A. Mircoli, and D. Potena, "A Negation Handling Technique for Sentiment Analysis," in 2016 International Conference on Collaboration Technologies and Systems (CTS), 2016, pp. 188–195.

[32] P. D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," n Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 417–424.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY