Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

Textual Sentiment Analysis using Lexicon Based Approaches

T. Nikil Prakash[1], Dr. A. Aloysius[2]
[1]Research Scholar, Department of computer science, St. Joseph's College, Thiruchirappalli, affiliated to Bharathidasan University, Thiruchirappalli.
[2]Assistant Professor, Department of computer science, St. Joseph's College, Thiruchirappalli, affiliated to Bharathidasan University, Thiruchirappalli.

Abstract

Sentiment analysis (SA) is a technique of textual data that uses Natural Language Processing (NLP) and Machine Learning (ML) to evaluate text automatically for the writer's feelings (positive, negative, and neutral). The lexicon-based approach is used to extracting sentiment from text and user reviews. In the sentiment analysis task, the sentiment lexicon, which offers sentiment polarity in terms, plays an important role. Most sentiment lexicons currently have only one polarity of sentiment for each word and disregard sentimental complexity. The problem of Sentiment Analysis was well studied and two main approaches were developed namely corpus-based and lexicon-based approaches. This paper discusses lexicon-based approaches to sentiment analysis. Contextual words, Acronyms, and emoticons are the major problems in sentiment analysis. The proposed techniques to improve the accuracy of sentiment analysis and also analyze the contextual words, acronyms, and emoticons.

Keywords: Sentiment Analysis, Lexicon-Based Approaches, Acronyms, Emoticons, Contextual Words, Natural Language Processing.

1. Introduction

Sentiment analysis is an evolving area of processing the natural language based on the interaction between humans and computers, extraction of information, and distillation of feelings from ever-increasing online social data. It includes recognizing the words or phrases indicating a positive, negative or neutral attitude in the underlying text. Sentiment analysis generally extracts various characteristics from structured or unstructured textual data and analyzes them to get thoughts, opinions, and feelings out of it. In this internet era, it is relatively easy to get the voice from customers or stakeholders through various channels, such as blogs, online forms, social media, customer service, and many more [1]. Typically, up to three different levels can be used in the sentiment classification namely (i) Document-level classification, (ii) Sentence-level classification, and (iii) Aspect-level classification. This research work performed a sentence-level sentiment classification in the research experiment. In several reviews, in a single product or service review, people express more than one opinion, typically distributed in various sentences.

In addition, lexicon-based methods are also popular in the study of sentiment, which takes into account the semantic orientation of words in a text and calculates sentiment. In this strategy, a dictionary of positive and negative terms is created where a sentiment value is assigned to each positive or negative word. These values are added to the text of the analysis, reduced to a bag of words, and mapped before with the dictionaries [2, 3].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

In this research work, to find a way to manage the difficulty of the data and offer a better result with better accuracy than the previous standalone lexicon-based methods. This research work addresses three critical issues of sentiment analysis.

- Some words have different meanings in different contexts.
- The complexity of acronyms, emoticons, and contextual words.
- Optimization of lexicon-based approach performance.

2. Literature Review

SamerMuthanaSarsam et al. [4] proposed suicide- and non-suicide-related emotion types on Twitter. They found the prevention of suicide risk and other mental-related disorders on Twitter. They found Suicide-related tweets which it contains a higher quantity of emotions that is fear, sadness, and negative. They found to perform better results for several existing approaches and classified suicide-related content on Twitter.

NurMaulidiahElfajr et al. [5] Focused on emoticon dictionary and weighting of an emoticon. They identified emotions in the sentiment by a sentence. They assumed that an emoticon state more observable emotions than words. They investigate the result was much better than the SentiWordNet process without an emoticon-based model.

HiranNandy et al [6] proposed filtering-based text sentiment analysis. They classified two main approaches for text sentiment analysis namely lexicon-based approaches and machine learning approaches. The proposed training data has been filtered built on six human emotions, and synonyms in English words. The proposed classifier model was a filtered training dataset. This model has given a significantly better performance than traditional machine learning-based models.

Alexandre Garcia et al [7] the protocol and a fine-grained opinion mining, based on a multimodal movie review dataset has been proposed. The token level method shows low inter-annotator agreements. The sentence-level achieved better values by relaxing the annotation granularity. The prediction of linear structure learns meaningful features level for the prediction of unusual labels.

Giuseppe Castellucci et al. [8] the contextual graph is described as a form in which messages can affect each other by taking into account links both intra-context and extra-context. The use of a Label Propagation algorithm confirms the positive effect of contextual information on social media in the Sentiment Analysis task. They improved the polarity value of words in the graph and explored the graph approaches in SA using the Modified Adsorption MAD algorithm. The problem of the proposed technique does not exactingly depend on the language of messages and effectiveness.

Milagros Fernandez-Gavilanes et al. [9] described an unsupervised SA strategy focused on semantic dependencies that are strengthened by SA of emoji creators' descriptions from Emojipedia. They developed a completely unsupervised emoji sentiment lexicon. This lexicon was once improved in a variety of ways that take advantage of the emotion distribution in

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

informal documents, such as emojis. They use a sentiment propagation algorithm to examine dependencies between lemmatized tagged terms, taking into account key linguistic phenomena such as intensification, negation, modification, and adversative and concessive relations.

YasirMehmood et al. [10] proposed an improved technique for lexicon-based sentiment analysis for social issues incorporating verbs with multi-level grammatical dependencies and improving the General Inquirer sentiment lexicon.. They compared ten online sentiment analysis tools to assess the efficacy of the proposed approach. Not only did the proposed solution outperform the online tools in terms of overall accuracy, but it also produced the best results for positive, negative, and neutral sentiment classifications. This study's findings are limited to its experimental setup of datasets gathered from a social issue: illegal immigration.

Khalid M.O. Nahar et al. [11] concentrated on the SA of Facebook Arabic comments for Jordanian telecommunications companies. The lexicon-based approach was used to determine the polarity of each of the provided Facebook comments. Data samples come from Jordanians commenting on a public issue concerning the services provided by Jordan's major telecommunications companies. The results of the evaluation of the Arabic sentiment lexicon were promising. They used the user-defined lexicon based on Jordanians most common Facebook posts and comments. They created a large dataset of unlabeled comments, the lexicon was used to label a set of Facebook comments.

3. Proposed Work

The proposed work is categorized into three levels. The first level is data collection and preprocessing, the second level is applying feature selection methods and the third level is feature classifications. Figure 1. Shows the proposed work framework. The proposed Senti_Con_Acro algorithm process provide several phases that is follows as:

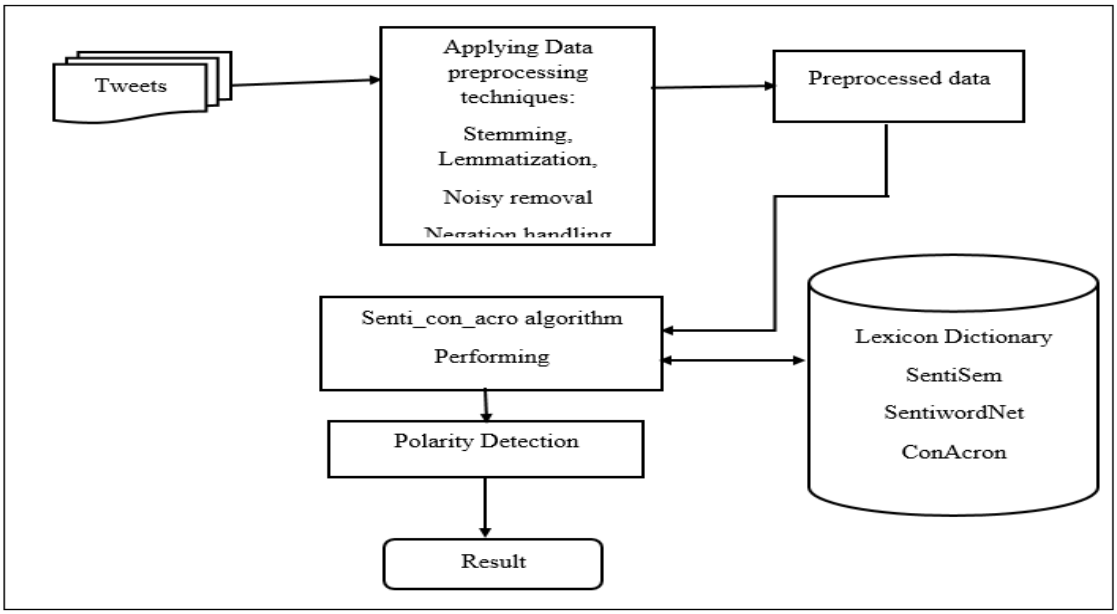

Figure 1: Framework for Senti_Con_Acron Algorithm

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

Level 1: Data Collection and Pre-processing

The Twitter dataset is collected in this research work. The collected data is stored and process to extract using Hadoop Distributed File System (HDFS) method [12]. Then the data is stored in .csv file format. The data preprocessing method is used to remove the stop words, noise data, images, audio, videos, etc. The data is cleaned then the data is process to the feature selection method.

Level 2: Feature Selection

The feature selection method is used for extracting features in the text. The contextual words, acronyms, and emoticons features are used for this proposed method. If the input text is given a different meaning in the sentence. Then the contextual word is used to extract the correct meaning of the sentiment for the given input text. The acronyms contain short words in the sentence and extract the original word in the sentence. The emoticon method is denoted as emojis that is provide happiness, anger, sadness, surprise, disgust, and neutrality. Unigram feature selection method is used for this proposed method. This method extracts the feature in single sentiments in the sentences. The SentiWordNet [13] dictionary classifies the sentiment feature in this proposed method. The text features are extracted from different ways that is follows as:

i. Unigram feature

In the proposed work Unigram featreus assume that the occurrence of each word is independent of its previous word. Hence each word becomes a gram (feature) here. For example:

"I", "have", "a", "lovely", "dog"

ii. Contextual words

The contextual words are called as different set of words or phrases. The proposed work identifies the user behavior or product performance in sentiment analysis. The contextual dictionary increasing contents constantly which provides unmatched opportunities to support decision-making processes and advocacy efforts.

iii. Emoticons

The emoticons are emoji's that identify the user behaviors and expressions. There are n number of emoji/ emoticons available in the emoticon dictionary. These emoji identify the positive and negative expression in sentiment analysis algorithms. The proposed work identify the emoticons and determining the emoticons are positive or negative value. Figure 1 displays sentiment polarity using emoticons.

iv. Acronyms

Abbreviations or acronyms are widely used in text materials to reduce space. The text in such areas consists of one to two sentences, or a few sentences such as text messages, social

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

media comments and blog posts. Customers may use or add new abbreviations or short word types, i.e. fast communication acronyms which rarely appear in regular or modern text, for these messages. Text as "TIA" for "Thank You in advance" is, for instance, common in these fields and for the machine the textual significance of the texts could hardly be accurately understood. The high-rate text adds new abbreviations that can impact the reliability of the emotional analysis. In order to solve this problem, abbreviations must be extracted and identified before the sentiment method is performed.

Level 3: Classification

After the feature is extracted in the sentiment then the polarity method classifies the test as positive, negative, and neutral. The proposed work features are evaluated using the confusion matrix method that is precision, recall, f-measure, and accuracy are used [14]. Table 1. Demonstrates the confusion matrix results in the proposed work.

Frequency Occurrence

Once the feature is extracted, they are used as input for supervised lexicon based approaches for further classification. Generally the frequency of occurrence of keyword is more suitable feature in overall sentiment analysis and not necessarily indicated by repeated use of keywords. Assumes that probability of each word occurring in a document is totally independent on word context as well as its position in the particular documents.

Brevity's Mandelbrot law equation is federated as the frequency of the sentiments which is measured as low rank and high-rank ratio and categorized through the deviancy of the power law. Brevity's Mandelbrot law check the ranking value if k >k0 is greater than the k0 value which gives the ranking is same in order the k value less than k0 the value is added as k0 + k

$$(Tt)\ \mathfrak{f}_k \propto (k + k)\text{-}b \qquad \text{eq.. (1)}$$

Where, f ← frequency of a word, k ← ranking of a word

Algorithm: Senti_Con_Acron

Input: DT ← Pre-processed data

Output: Classes of sentiments like positive, negative, neutral

For each $item \in$ DTdo

dt ← $item$ [$'text'$]

Apply Unigram feature model

Count each sentiment

If 'dt' item ['Neighboring Words']

Replace correct contextual words

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

Else if 'dt' item ['Acronyms']

Replace the correct sentiment words

Else if 'dt' item ['Emoticons']

Replace the correct emoticon sentiments

End for

Calculate the polarity value like positive, negative, and neutral

Overall Results: precision, recall, f-measure, accuracy.

Find frequency rank: (dt) $\mathfrak{F}_k \infty (k + k)$-b

4. Results

Social media (i.e. Twitter) dataset is collected from this proposed work. The data is cleaned and processed using the preprocessing algorithm. The proposed work focuses on contextual words, acronyms, and emoticon sentiments. The Senti_Con_Acron model improves the better accuracy of existing models. The Senti_Con_Acron model efficiency is compared to other existing methods. Table. 2. Demonstrates the comparison results of existing work with proposed work.

Table. 1. Proposed work results for Precision, Recall, F-Measure, Accuracy

	Precision (%)	**Recall (%)**	**F-Measure (%)**	**Accuracy (%)**
Proposed Result	76.75	73.52	75.14	80.13

Table. 2. Comparison results for proposed work

Authors	**Results**
Ahmad Aloqaily et al. [15]	68
M. Edison et al. [16]	68.75
VallikannuRamanathan et al. [17]	76
SeydehAkramSaadatNeshan et al. [18]	76.3
Proposed Work	**80.13**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

5. Conclusion

Sentiment analysis has developed one of the energetic methods in creating commercial decisions as it directly comprises the customer group. Despite the current evolution in this research area, there are still many challenges as human attitudes and write up in the form of review is difficult and uncertain. The proposed work is to classify contextual words, acronyms, and emoticons using lexicon-based approaches. The SentiWordNet dictionary is used to identify the sentiment and increase the performance of lexicon based approaches. The proposed model improves the better accuracy results compared to the existing works. In the future, we can apply machine learning techniques to improve performance.

References

[1]. TityaEng, MdRashed Ibn Nawab, KaziMdShahiduzzaman "Improving Accuracy of The Sentence-Level Lexicon-Based Sentiment Analysis Using Machine Learning" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, DOI: https://doi.org/10.32628/CSEIT21717, Vol-7 (1), 57-68, 2021.

[2]. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," Secure. Inform, Vol- 4, no. 1, 1-9, 2015.

[3]. Gupta, J. Pruthi, and N. Sahu, "Sentiment Analysis of Tweets using Machine Learning Approach," Int. J. Comput. Sci. Mob. Comput, Vol- 6, no. 4, pp. 444–458, 2017.

[4]. SamerMuthanaSarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Waleed Alnumay Andrew Paul Smith, "A lexicon-based approach to detecting suicide-related messages on Twitter" Biomedical Signal Processing and Control, DOI: https://doi.org/10.1016/j.bspc.2020.102355, Vol- 65, 2021.

[5]. NurMaulidiahElfajr, RiyanantoSarno "Sentiment Analysis using Weighted Emoticons and SentiWordNet for the Indonesian Language" International Seminar on Application for Technology of Information and Communication, pp. 234-238, 2018.

[6]. HiranNandyEmailauthorRajeswari Sridhar, "Filtering-Based Text Sentiment Analysis for Twitter Dataset" Advances in Artificial Intelligence and Data Engineering, pp 1035-1046, 2020.

[7]. Alexandre Garcia, Slim Essid, Florence d Alche-Buc, Chloe Clavel "A multimodal movie review corpus for fine-grained opinion mining" arXiv: 1902.10102v1 [cs.MM], 1-14, 2019.

[8]. Giuseppe Castellucci, Danilo Croce and Roberto Basili "A Graph-based Model of Contextual Information in Sentiment Analysis over Twitter" Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015, DOI: 10.4000/books.aaccademia.1332, 72-76, 2016.

[9]. Milagros Fernandez-Gavilanes, Jonathan Juncal-Martínez, Silvia García-Méndez, Enrique Costa-Montenegro, Francisco Javier González-Castano "Creating emoji lexica from unsupervised sentiment analysis of their descriptions" Expert Systems With Applications, Vol. 103, 74–91, 2018.

[10]. YasirMehmood and VimalaBalakrishnan "An enhanced lexicon-based approach for sentiment analysis: a case study on illegal immigration" Online Information Review, Emerald Publishing Limited, DOI 10.1108/OIR-10-2018-0295 , pp. 1468-4527,2020.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9878 - 9885
Received 05 March 2021; Accepted 01 April 2021.

[11]. K. M. O. Nahar and A. Jaradat, M. S, F. "Sentiment Analysis And Classification Of Arab Jordanian Facebook Comments For Jordanian Telecom Companies Using Lexicon-Based Approach And Machine Learning" Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 06, No. 03, 2020.

[12]. Dipty Sharma "Study of Sentiment Analysis Using Hadoop" Big Data Analytics, AISC, Vol- 654 pp. 363-376, 2018.

[13]. MaiteTaboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede "Lexicon-BasedMethods for Sentiment Analysis" Association for Computational Linguistics, Volume 37, no 2, 2011.

[14]. Alexander Pak, Patrick Paroubek "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" Proceedings of the International Conference on Language Resources and Evaluation, Researchgate, 1320-1326, 2013.

[15]. Ahmad Aloqaily, Malak Al-Hassan, Kamal Salah, BasimaElshqeirat, MontahaAlmashagbah, "Sentiment Analysis for Arabic Tweets Datasets: Lexicon-Based and Machine Learningapproaches" Journal of Theoretical and Applied Information Technology, Volume 98. No 04, PP. 612- 623, 2020.

[16]. M. Edison and A. Aloysius "Lexicon based Acronyms and Emoticons Classification of Sentiment Analysis (SA) on Big Data" International Journal of Database Theory and Applications, DOI: http://dx.doi.org/10.14257/ijdta.2017.10.7.04, Volume.10, No.7, PP. 41-54 2017.

[17]. VallikannuRamanathan and Meyyappan T "Prediction of Individual's Character in Social Media Using Contextual Semantic Sentiment Analysis" Springer Science, DOI: https://doi.org/10.1007/s11036-019-01388-3, PP. 1-15, 2019.

[18]. SeydehAkramSaadatNeshan and Reza Akbari "A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis" IEEE, PP. 1-7, 2020.

[19]. T. Nikil Prakash and A. Aloysius, "Lexicon Based Sentiment Analysis (LBSA) to Improve the Accuracy of Acronyms, Emoticons, and Contextual Words", Society of Statistics, Computer and Applications, (Accepted- Forthcoming), 2021.

[20]. T. Nikil Prakash, A. Aloysius "Applications, Approaches, and Challenges in Sentiment Analysis (AACSA)", International Research Journal of Modernization in Engineering Technology and Science, ISSN: 2582-5208, Vol. 2, Issue 7, pp. 910-915, 2020.