Pakistan Journal of Science, 71 (4 Suppl.): 2019, Page 230-234
ISSN: 2411-0930

SENTIMENT ANALYSIS BASED ON ABBREVIATIONS IN TEXT

S. Malka*, S. Jan and I. A. Shah

Department of Computer Software Engineering
University of Engineering and Technology, Mardan, Pakistan
*Corresponding author's E-mail: malka.shah12@gmail.com

ABSTRACT: Microblogging has become one of the prevalent sources of communication. A large volume of data is available online in the form of text across the digital world on social media websites, blogs, news, articles, digital libraries, comments and reviews. The aim is to organize and manage data and obtain valuable information from it. Data volume has grown exponentially over the past few years and it has become impossible for humans to retrieve the required information easily. For this purpose, various text mining techniques have been proposed and used including sentiment analysis. Websites have become a rich source of data and the increasing use of acronyms in text has emerged as a new challenge which affects the results of the sentiment analysis. The case worsens in the informal text like opinion and reviews. This paper presents a novel idea of enhancing the accuracy of sentiment analysis by dealing with acronyms in a particular domain. The proposed system was used for many domains and it shows significant improvement in overall accuracy of the text analysis. The proposed system was evaluated using metrics of information retrieval, i.e., Precision, Recall, and F-Measure. The experimental results show that the system achieves 57% Precision, 78% Recall and 64% F-Measure before mapping the long form of the acronyms. Moreover, the system performance is increased by achieving 89% Precision, 91% Recall and 90% F-Measure after mapping the long-form of acronym into the text.

Keywords: Data Mining, Text Mining, Acronyms Identification, Abbreviation Mapping, Sentiment Analysis.

INTRODUCTION

A large volume of data is available online in the form text across the digital world in the form of social media websites, blogs, news, articles, digital libraries, comments and reviews. It has become a challenging task to analyze and organize such large collection of data for extracting the valuable information from it. The process of identifying valuable information and knowledge from the unstructured textual data is an emerging field and is referred to as text mining (Sajid *et al.*, 2017). It has the wide range of applications comprising of competitive intelligence, fraud detection and security, analyzing online suspicious activities, finding point of views in the news articles and blogs, customer relationship management, online market trend analysis and product reviews (Patel and Mistry, 2015). Sentiment analysis is the computational study of people's attitude and emotion toward an object. During sentiment analysis, the sentiments are identified and then analyzed (Ravi and Ravi, 2015). It is a type of natural language processing that tracks out the attitude of the individuals toward an event or object. Sentiment analysis involves developing a system that captures and analyses opinion about the event or an object in social media comments, blog posts or review feedbacks. Sentiment analysis is highly useful in many domains (Balazs and Velásquez, 2016). The text in such areas such text messages, social media comments and blog posts, consist of one to two sentences or few phrases. Also, when constituting such messages, user may use or can add new abbreviations or short forms of the words, i.e., acronyms for quick communication, that rarely appear in standard or conventional text. For example, text like "TIA" for "Thank You in Advance" is popular in such domains and for the machine it is hard to accurately recognize semantic meanings of these texts; which is one of the several challenges to the sentiment analysis. New abbreviations are added to the text with high rate which can affect the accuracy of the sentiment analysis (Giachanou and Crestani, 2016). To solve this problem, the extraction of abbreviations and their definition is required before applying the sentiment analysis. In this study an algorithm is developed uses lexicon-based approach to find the sentiment orientation of the text after identification and replacement of acronyms with its long form. The approach uses contextual semantic orientation of each word in text to identify the sentiment of the text.

The rest of the paper is organized as follows: Section II describes the related research work done in different text mining techniques. Section III presents the proposed research work and implementation of the system. Section IV presents results, performance analysis, evaluation, and comparison of the result before and after mapping abbreviation in the text. Finally,

Section V concludes the paper with hints for the future work.

MATERIALS AND METHODS

The identification of opinion in text data is a well-known field of research. It has been considered for studying the reviews in different domains, i.e., for product evaluation by reviewing comments of the customers (Avanço and Nunes, 2014), for faculty evaluation in education field, for improving health service (Greaves *et al.*, 2013), for identification tourist destinations (Toral *et al.*, 2018) and in many other fields. In (Liu, 2012) the Sentiment analysis has been covered from different prospective. Opinion extraction from the text can be done using machine learning or lexicon-based Approach. In this study, lexicon-based approach is used to find the opinion in a text.

Machine learning approach: Machine learning approach trains classifier using labelled training dataset. This approach requires large training data for the classification of the text (Hailong *et al.*, 2014). In the work of (Wawre and Deshmukh, 2016) Machine learning is used to find the opinions in the movie reviews using two method of machine learning i.e., supervised machine learning approach and naïve Bayesian approach and concluded that naïve Bayesian approach outperforms SVM.

Lexicon based approach: This approach is based on consideration that the contextual sentiment orientation is the collection of each word or phrase's sentiment orientation. Which can be generated using two approaches, i.e., corpus based approach and dictionary based approach (Gitari *et al.*, 2015). The lexicon-based approach was used to identify the sentiment of the text using semantic orientation calculator (SO-CAL) for which the dictionaries of words labelled with their semantic orientation, i.e., positive and negative labels. SO-CAL was used for classification of text to identify the opinion in the text (Taboada *et al.*, 2011). The extraction of sentiment is more challenging on microblogs as compared to other textual data. These challenges are due to presence of spelling mistakes, short length messages, informal words, slangs, emoticon, and acronyms. In the work of (Khalifa and Omar, 2014), the same approach of lexicon based with naïve Bayes classifier is used to identify the sentiment toward the services of the product in Arabic text. They first executed the lexicon approach by substituting words with its synonyms using domain dictionary and the classification into positive and negative was performed using naïve Bayes classifier. The work of (Hu and Liu, 2004) performed the sentiment of twitter data using lexicon-based approach to classify tweets as positive, negative or neutral. They showed that the size of lexicon and its coverage directly affects the accuracy.

A step by step methodology for the lexicon-based approach is applied to find the sentiment of the data that contains the acronyms, i.e., short forms of the words. We show that the accuracy of the sentiment analysis can be improved by mapping long form of the acronym in the text. The first step in finding the sentiment analysis is the text pre-processing called data preparation phase as described in the next sub-section:

Data Pre-Processing: The steps of pre-processing are applied before applying the actual procedure of sentiment finding. The purpose of this step is the normalization of the text into a suitable format to extract the sentiments. Which improves the overall accuracy of the classification of the text (Uysal and Gunal, 2014). The data preparation phase includes the following steps.

Removal of HTML and XML tags: Beautifulsoup Python library has been used to extract the data from HTML and XML tags. It also provides an interface for webpage crawling.

Remove Square Brackets: The second step of the preprocessing was the removal of square brackets from the text. For this purpose, built in function RE (regular expression operation) is used

Denoising Text: The third step of the preprocessing was denoising of the text that includes removing of tags and metadata using Beautifulsoup library

Tokenization or Bag-of-Words Creation: In this step, the strings of text are broken down into token, i.e., words, phrases, symbols by using NLTK (natural language tool kit). In this step of the preprocessing, the white spaces between the words are also removed from the text.

Lowering the Case of the Letters: After normalization all the tokens are converted into lower case using NLTK library from the list of tokenized words.

Removing Punctuation: This step removes punctuation from list of tokenized words.

Removing Numbers: In identifying sentiment of the data, numbers are of no use thus this step replaces all integer occurrences in list of tokenized words with textual representation.

Stop-words Removal: In a sentence the connecting function is performed by stop words. These words include words such as preposition, articles such as; is, the, at, and on. This step removes the stop words from the list of tokenized words using NLTK library.

Stemming: during this process the morphological variants of a root/base word is produced. This step uses NLTK library to stems the derivation of word into its own

roots. For example, the words; loving, loved, and loveable could be reduced into love.

Negations Handling: In this step, the negation of words is handled. In this process the sentiment of the text is converted from negative to positive and from positive to negative using special words such as not, no etc. the negation of words plays an important role in identifying the sentiment of the text. It can change the whole sentiment of the text. For example, good is identified as positive whereas not good is identified as negative.

Classification of Text: Opinion in the text is identified after the pre-processing step. In this step the lexicon-based method is used to determine the sentiment prediction of the text. In this approach, a polarity score with each word is associated which shows the semantic orientation of the words, i.e., positive or negative. The contextual semantic orientation of each word in text is used to identify the sentiment of the overall text. The words in the review (text) were compared with the words in the list labelled with their semantic orientation. Based on which the overall sentiment of the text is calculated, and the text is classified as positive or negative. The list we used in our study consisted of 6800 sentiment words with the polarity score associated with each word and is compiled by in their work. The opinion of the text is identified according to the existing list of positive and negative words.

Identification of Acronyms: The process of extraction of acronyms, slangs and their definition mapping were performed after identifying the sentiment of the reviews. The system was first evaluated before the identification of acronyms for the baseline result using the metrics of information retrieval, i.e., Precision, Recall and F-Measure followed by the extraction and replacement of acronym with its definition. During this step, the first task was the identification of acronyms and slangs in the text. The second task performed was identifying the long form of these words. For this purpose, we crawled different websites in order to get the list of most commonly used acronyms. Table 1 showing the snap of some of the acronyms and their definitions that are maintained in acronym dictionary, which are replaced with acronyms present in the text after the identification of these acronyms. Whereas Fig.1 and Fig.2 show the overall proposed work.

Figure 1 shows the first part of our proposed work i.e., steps performed in order to find the sentiment orientation of the text containing acronyms in its original form.

Figure 2 shows the second part of our proposed work i.e., steps performed to find the sentiment orientation of the text after handling the acronyms present in the text by mapping it to long form i.e., definitions

Table 1. Acronyms Dictionary.

Acronyms	Long Form
Lol	laughing out loudly
tia	thank you in advance
gr8	great
rip	rest in peace
yolo	you live only once
Asap	as soon as possible
Bff	best friends forever

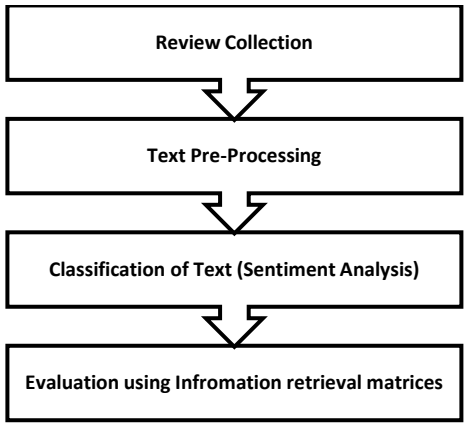

Figure-1: Part first of the Proposed work

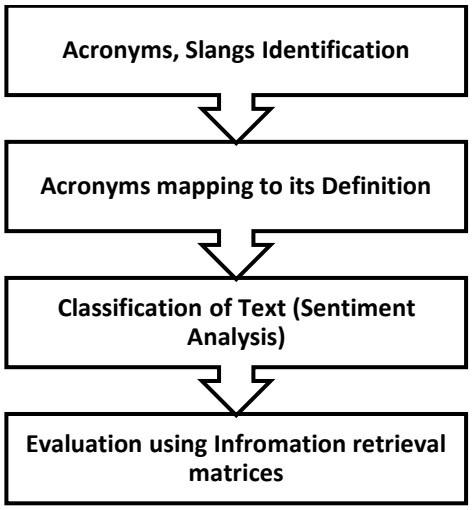

Figure-2: Part second of the Proposed work

RESULT AND DISCUSSION

The proposed approach has been applied to the users' reviews on different applications of Google Play store. The reviews contained acronyms, slangs etc. The dataset was collected and maintained by Kaggle which is publicly available at (Greaves *et al.*, 2013). First, we identified the sentiment of the text and then evaluate the result using information retrieval metrics such as Precision, Recall and F-Measure. After which the

acronyms, slangs and informal words present in text were mapped to its definitions, i.e., long form. Then we identified the sentiment of the text after mapping. The result obtained after mapping the definition of acronyms present in the text is evaluated against the result of the sentiment of the text before mapping the long form.

Identification of Sentiment: The first step in the process was the identification of the sentiment of the text. Table 2 shows few examples of the reviews with their semantic orientation. Whereas Fig 2. shows the system's console screen presenting the sentiment of the review.

Table 2. Review with Semantic Orientation.

Review	Sentiment
Gud. Because.. YouTube video recording. front camera best candy camera	Positive
Faltu plz waste ur time	Negative
I downloading game right I good #YOLO	Positive

Performance Parameter: The sentiment prediction performance has been evaluated using evaluation metrics, i.e., Precision, Recall, and F-Measure. Precision is the percentage of relevant words to the total number of words identified by the system. Recall is the percentage of the number of relevant words identified by the system to the total number of relevant words. It shows the result in the form of Precision, Recall, F-Measure of the sentiment prediction of the reviews before mapping the definition of the acronyms present in the text. The Precision, Recall, F-Measure achieved is 0.57, 0.78, 0.64 respectively. Whereas Fig. 4. shows the result of the sentiment prediction of the text after mapping the definitions of acronyms in the text. The Precision, Recall, F-Measure achieved is 0.89, 0.90, 0.91 respectively.

Results Comparison: Figure 5 shows the graphical comparison of the results of sentiment prediction before mapping the long form of acronyms and after mapping the long form of acronyms to the text. It has been observed through information retrieval matrices, i.e., Precision, Recall and F-Measure, that the sentiment prediction increases after mapping the long form of acronyms into the text. In the presence of the long form of the acronyms, the Precision increases from 0.57 to 0.89, Recall increases from 0.78 to 0.9 and F-Measure increases from 0.64 to 0.91 of the sentiment predictions of the text. This shows that the technique of mapping acronyms to its long form before finding sentiment predication of the text outperforms than the sentiment predication of the text having acronyms, in its original form.

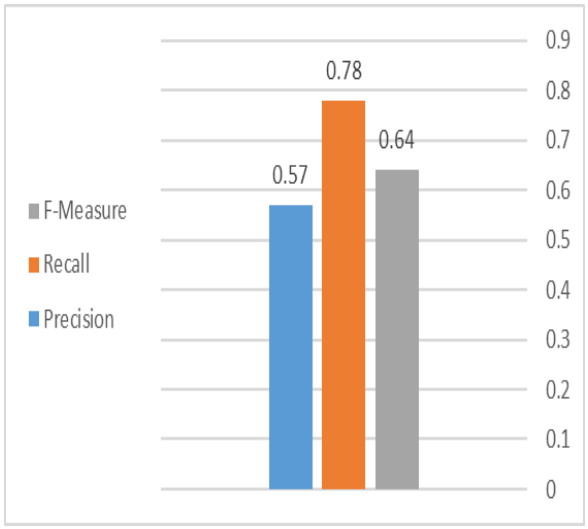

Figure-4: Results Before Mapping definitions

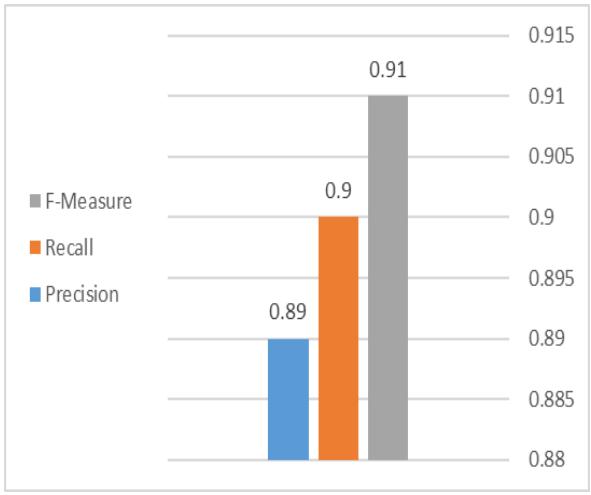

Figure-5: Results after Mapping Definitions

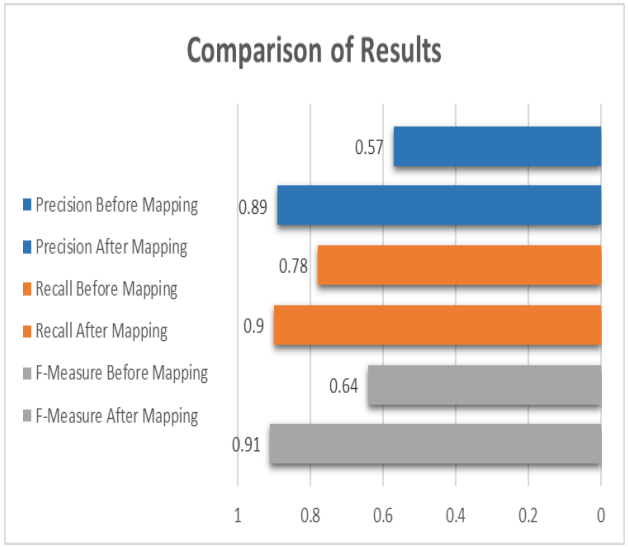

Figure-6: Results after Mapping Definitions

The system either ignores these acronyms or treat them as a neutral word. Figure 6 shows the snap of the review whose semantic orientation changed after mapping acronyms to its long form.

Analysis of Error: Table 3 shows the example of incorrectly classified review. This incorrect result is due to the presence of ambiguous word. The proposed system does not detect the emotions of the words, e.g., in the phrase "it was a damn good", "damn" has been identified as a negative word. Where it is not used as a negative word but is a figure of speech called hyperbole which is used for the exaggeration of the word "good".

Table 3: Incorrect semantic Orientation.

Review	Semantic Orientation
It's osm n filters damn gud	Negative
Freezes scrolling bit. Every damn time.	Negative
i nearly died laughing	Neutral

Conclusion: This work has showed that the presence of acronyms in the text can compromise the performance of the system in identification of the semantic orientation of the text. The system ignores these words i.e. acronyms or labelled these words as neutral which effects the overall sentiment of the text. It has been observed that the mapping of these acronyms, slangs or informal words to its definitions i.e. long form before identification of the sentiment prediction of the text can increase the performance of the system. Since some of the words such as hyperbole figure of speech are interpreted incorrectly. This can be improved in the future by incorporating the hyperbole detection in the system.

REFERENCES

Avanço, L.V. and M.D.G.V. Nunes (2014). Lexicon-based sentiment analysis for reviews of products in Brazilian Portuguese. In *2014 Brazilian Conference on Intelligent Systems* (pp. 277-281). IEEE.

Balazs, J.A. and J.D. Velásquez (2016). Opinion mining and information fusion: a survey. *Information Fusion, 27*, 95-110.

Giachanou, A. and F. Crestani (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR), 49*(2), 28.

Gitari, N.D., Z. Zuping, H. Damien and J. Long (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215-230.

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi and L. Donaldson (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research, 15*(11), e239.

Hailong, Z., G. Wenyan and J. Bo (2014). *Machine learning and lexicon based methods for sentiment classification: A survey.* Paper presented at the 2014 11th Web Information System and Application Conference.

Hu, M. and B. Liu (2004). *Mining and summarizing customer reviews.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Khalifa, K. and N. Omar (2014). A Hybrid method using Lexicon-based Approach and Naive Bayes Classifier for Arabic Opinion Question Answering. *JCS, 10*(10), 1961-1968.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Patel, P. and K. Mistry (2015). A review: Text classification on social media data. *IOSR Journal of Computer Engineering, 17*(1), 80-84.

Ravi, K. and V. Ravi (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46.

Sajid, A., S. Jan and I.A. Shah (2017). *Automatic Topic Modeling for Single Document Short Texts.* Paper presented at the 2017 International Conference on Frontiers of Information Technology (FIT).

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Toral, S., M. Martínez-Torres and M. Gonzalez-Rodriguez (2018). Identification of the unique attributes of tourist destinations from online reviews. *Journal of Travel Research, 57*(7), 908-919.

Uysal, A.K. and S. Gunal (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), 104-112.

Wawre, S.V. and S.N. Deshmukh (2016). Sentiment classification using machine learning techniques. *International Journal of Science and Research (IJSR), 5*(4), 819-821.