Domain Specific Lexicon Generation through Sentiment Analysis

https://doi.org/10.3991/ijet.v15i09.13109

Kamran Shaukat
The University of Newcastle, Newcastle, Australia

Ibrahim A. Hameed (✉)
Norwegian University of Science and Technology, Trondheim, Norway
ibib@ntnu.no

Suhuai Luo
The University of Newcastle, Newcastle, Australia

Imran Javed, Farhat Iqbal, Amber Faisal, Rabia Masood, Ayesha Usman,
Usman Shaukat, Rosheen Hassan, Aliya Younas, Shamsair Ali, Ghazif Adeem
University of the Punjab, Pakistan

Abstract—Sentiment analysis (SA) is used to extract opinions from a huge amount of data, and these opinions are comprised of multiple words. Some words have different semantic meanings in different fields, and we call them domain-specific (DS) words. A domain is defined as a special area in which a collection of queries about a specific topic are held when user do queries in the data regarding the domain appear. But Single word can be interpreted in many ways based on its context-dependency. Demonstrate each word under its domain is extremely important because their meanings differ from each other so much in different domains that a word meaning from A in one context can change into Z in another context or domain. The purpose of this research is to discover the correct sentiment in the message or comment and evaluate it either it is positive, negative or neutral. We collected tweets dataset from different domains and analyzed it to extract words that have a different definition in those specific domains as if they are used in other fields of life. They would be defined differently. We analyzed 52115 words for finding their DS meaning in seven different domains. Polarity had been given to words of the dataset according to their domains and based on this polarity, they have been recognized as positive negative and neutral and evaluated as domain-specific words. The automatic way is used to extract the words of the domain as we integrated and afterwards the comparison to identify that either this word differs from other words as far as domain is concerned. This research contribution is a prototype that processes your data and extracts their domain-specific words automatically. This research improved the knowledge about the context-dependency and found the core-specific meanings of words in multiple fields.

Keywords—Domain-specific; General-language dictionary; lexicon; polarity; Sentiment analysis

1 Introduction

Sentiment analysis is a somewhat natural language dealing with following the perspective of the all-inclusive community about a particular thing or point. Sentiment analysis, which is moreover known as opinion mining, incorporates into building a system to accumulate and take a gander at suppositions about items made in the blog sections, comments, overviews or tweets. Sentiment analysis can be useful in a couple of ways. For example, in showcasing it helps in judging the accomplishment of an ad campaign or new thing dispatch, make sense of which variations of a thing or administration are notable and even recognize which socioeconomic like or particular abhorrence elements [1].

Generally, the Sentiment analysis structures have focused on specific areas using space-specific corpora as preparing information for the machine learning calculations that organize a data message as either the positive or the negative. Diverse structures are vocabulary-based, where assessment bearing words and articulations are accumulated and a short time later scanned for amid examination to concoct a specific sentiment index.

SA is classified into three levels, i.e. aspect level, document level and sentence level. The aspect level of the sentiment analysis would allude to supposition related to parts of the substance being talked about. This takes into consideration a more detailed investigation that uses a greater amount of the information given by the textual review [2].

With the quick advancement of Web innovation, individuals are getting increasingly data from the web. Instructions given to the individuals about concerned data from tremendous measure of the web information, sentiment polarity analysis has gained more importance [3]. SA polarity is defined as the scoring positive and negative effect of a text or comment [4]. In the previous researches, the most common way for classifying the text messages such as positive, negative and neutral keep up to the subject [5-8]. For the improvement of individual sentiments words, sentiment dictionaries are used. The name of the sentiment dictionaries is SentiWordNet and WordNet [9]. SentiWordNet is a domain-independent openly accessible method for utilization of SA. SentiWordNet is extracted from WordNet by giving positive or negative scores [10].

Today due to the quick advancement of the web technology, the people are sending and receiving data from the web. To answer all the questions about web technology; sentiment polarity analysis is becoming increasingly dominant. Different authors have done their job to find out the polarity of the things; for example, in financial news comments or messages used negative and positive polarity. In others, exploring domain adjustment for sentiment concentrating on online surveys for various sorts of items and assigning them polarity [11-17]. In DS SA, there is a basic need to build a DS corpus. When a user does queries, the data lies in the domain [18, 19].

For a domain-independent approach, three types of lexical ways are used such as SentiWordNet, Senticnet and SentiSlangNet. We used Senticnet for comparison. Senticnet is an openly accessible semantic means which consists of frequent utilization of polarity ideas [20-22]. Senticnet is also known as an openly accessible way that is built by methods for evaluation or figuring the sentic [23, 24]. Twitter is a mainstream microblogging service where clients make status messages called "tweets". Sometimes these tweets express sentiments about various themes. Twitter messages are likewise utilized as a source of data for sentiment classification. In SA twitter data is used for peculiar plans [25, 26].

With the use of the SA technique, we can analyze which renditions of an item or service are prominent and even distinguish which demographics like or dislike specific features and also it is a bridge between two domains. The purpose of this research is to identify words that are interpreted differently among various fields; that will help in finding the core-specific meaning of a word and also improve the performance of different domains.

2 Literature Review

SA is considered a big data task. A scalable lexicon-based approach for squeezing sentiment using emotion and hashtag is introduced for good performance in accuracy and speed in [27]. Soni et al. [28] divided the work into two parts, in the first part Hidden Markov model is proposed and in the second part test and shows the comment to analyze the opinions of the person about the product. POS sequence has been proposed as an attribute for investigation of pattern or the word combination of tweets in two different domains of the sentiment analysis. One is subjectively, and the other is polarity. To get the most perfect in uncovering sentence pattern Three Forms of POS sequence is used, i.e. the pattern of 2-tags, 3-tags, and the 5-tags [29]. Lu [30] proposed a novel approach of semi-supervised learning for MSA (Microblog Sentiment Analysis). A graph-based semi-supervised classifier is built to make use of microblog relations. This method connects tagged and untagged data through microblog relations. Arora et al. [31] took SA and collected information used by the business for improving the product through it. The author has compared the customer's sentiments given via tweeter on five smartphone brands which include battery timing and other operating system using one week of tweeter data.

It's hard to describe real-life complications with low-level features. Do et al. [32] proposed a framework named emotion prediction framework using the high-level features of contextual. This framework identifies the emotion of the situation using the high-level features of contextual. Pawar et al. [33] describes the level, approaches, and methodologies of doing sentiment analysis and provide a feature to squeeze from the text and application. This paper provides an overview of tweet extraction, their preprocessing and sentiment analysis.

Medhat et al. [34] have categorized various SA techniques. The fields of SA, which include TL (Transfer Learning), ED (Emotion Detection), and BR (Building Resource), are discussed. ML algorithm is used to solve the SC problems. NLP (Natural Language Processing) tool is used to reinforce the SA process, which has attracted researchers. Guzman et al. [35] used natural processing techniques that recognize the fine-grained app feature in the review. Collocation finding problems are combined which are lexical sentiment analysis, and producing summaries of recall up to 73% (51% average) and a precision up to 91% (59% average) for the topic modelling.

Bansal [36] described PCAP (Packet Capturing) tools that are used for the protection of private data and network infrastructure against attacks. PACP and related tools are available as open-source software. Python programming language is used to handle any type of data stream. To Build a DS lexicon Park et al. [37] appraised the extracted lexicon against SentiwordNet and existing algorithm. Active learners are used to increasing the F1-score in classification and increased similarity of sentiment lexicon. This proposed system is generally used for N-gram cases.

Khuc [38] used two proposed methods which are graph propagation and Labeled TNG methods. TNG method is better to use for n-grams. It gives higher accuracy than graph propagation. And the key focus of this work was automatically generating polarity lexicons for sentiment analysis on social networks. Agarwal et al. [39] explored the effect of three factors that are used to discover the all-inclusive text sentiment. Firstly he explored DS Ontology using ConceptNet based ontology. Secondly, he described the importance of features. Lastly, he explored the contextual information.

SA of online social networks Trung et al. Authors in [40] have proposed a fuzzy propagation model for opinion mining. Tweetscope; a practical system is made which gathers and evaluates the customer tweets. Hussain et al. [41] worked on feature categorization and feature extraction using NLP techniques to find out the feature about the product review etc. in feature-based SA. There are two types of feature extraction. The first type is explicit feature extraction (FE) and 2nd is implicit FE. Abundant work is done on explicit FE instead of Implicit FE and feature clustering using NLP technique in feature-based SA.

Kontopoulos et al. [42] proposed ontology-based techniques for twitter's tweets. The basic working of these techniques is to assign a sentiment score. The first step of this technique is to find the subject of the tweet then break them into parts according to the subject defined. There are many approaches based on machine learning that are used to perform SA, but the disadvantage of these approaches is that they take each tweet as one equable statement and assign score as a whole. Choi et al. [43] proposed a domain-specific SA system which has two parts. The first part is a context feature generation. The second part is the DS sentiment classifier learning. The bootstrapping method is applied to make the Domain context classifier using CF. This method gives benefit for extracting contextual clues in the news domain and increases the performance of sentiment classification.

Authors in [44] proposed a mining model to extract the semantic relation to discover the actual sentiment present in the conversation and to find the knowledge level and learning style of the students in the e-learning environment. The under-observation conversations consisted of the rich amount of domain specific data. Doman ontology was used to extract the information from unstructured chats. The learning process was used only for a small group of students and for a limited number of subjects.

Authors in [45] used a sentiment classification tool LDA (Latent Dirichlet Allocation) to measure the feelings of students about some specific subjects and topics by classifying documents according to the sentiment present in the chats of the students. The main goal of the work was to help the teachers to enhance the e-learning environment according to the moods of the learners.

Authors in [46] conducted different experiments to find the effectiveness of the conventional method of vocabulary learning and computer corpus-based vocabulary learning to improve the English vocabulary learning of students. The result showed the computer corpus-based learning method performed efficiently than the conventional learning method. However, students need to use computer corpus-based vocabulary under the guidance of the teachers as the majority of the students are not familiar with this method.

Authors in [47] proposed a new conceptual model of interests and sentiments, also used an already existing computational model to discover the significance of sentiment analysis in e-learning. A real-life e-learning environment was experienced to apply both specified models. The study was based on self-reported sentiments of university students. The authors emphasized that sentiment analysis, combined with the lexical analysis, can serve as an implicit method to measure the learner's sentiments. More learning environments are needed to be observed to validate the already obtained findings.

3 Methodology

This section explains the overall methodology of our proposed work. Fig 1 shows the graphical representation of the proposed work.

3.1 Dataset collection

Different researchers used the different features to mine sentiment analysis on domain-specific criteria. This research is going to extract words that have different meanings in different domains so that they should not be treated as same in all the fields.

Table 1. Dataset Summary

Dataset Attributes	
Total number of words	52115
Number of domains	7
Politics dataset words	7744
Terrorism dataset words	7717
Life dataset words	7040
Science dataset words	7449
Sports dataset words	7734
Gossips dataset words	7091
Movies dataset words	7340

The dataset used in this research is extracted from twitter using R-studio [48] software. All the work in R-studio is done by using r language. The data about nine following domains have been collected.

Politics Movies Terrorism Sports
Science Life Gossip

All the data collection of these domains has been collected under the impression of the social domain. The social domain is one giant domain of multiple sub-domains and that's the reason the dataset has been collected from twitter (a social networking site).

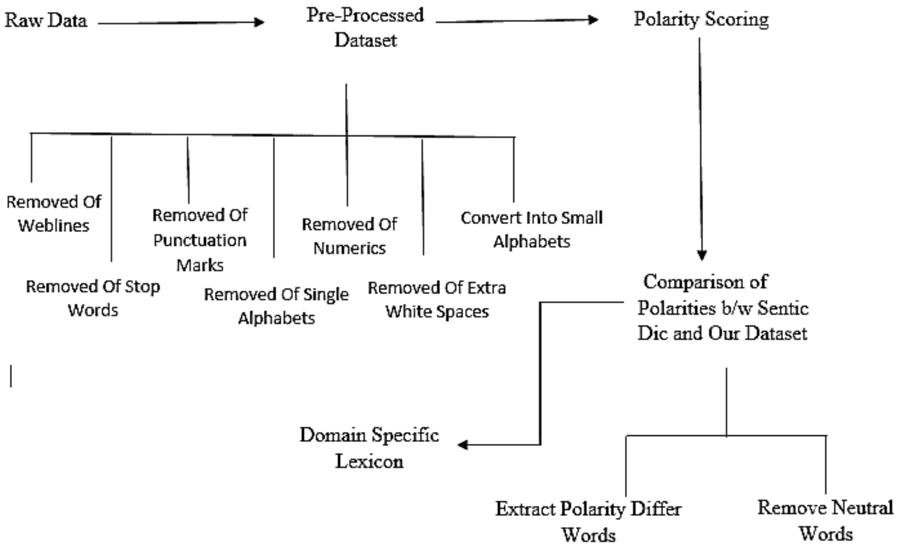
Raw Data
Pre-Processed Dataset
Polarity Scoring
Removed Of Weblines
Removed Of Punctuation Marks
Removed Of Numerics
Convert Into Small Alphabets
Removed Of Stop Words
Removed Of Single Alphabets
Removed Of Extra White Spaces
Comparison of Polarities b/w Sentic Dic and Our Dataset
Domain Specific Lexicon
Extract Polarity Differ Words
Remove Neutral Words

Fig. 1. Proposed Methodology

3.2 Preprocessing

We performed the following seven steps to clean the collected dataset, which are as follows:

HTML Removal: All the web links either at the end of tweets or links, someone given as a tweet are removed manually. By doing so ,we got rid of fuzzy and improper words used to name the websites and also for this kind of research web links are considered raw data that does not need to convert into information.

Punctuation Marks Removal: All the punctuation marks such as comma, colon and question mark must be removed from the dataset as these marks make the data noisy. This step is done on our data in R-studio.

Removal of numerical: In our research, we only need words to process, so all the digits in data are removed, so our work became easy as we did not get to deal with the complicated numerical. We also perform this step in R-studio.

Conversion of Data into Small Alphabets: Capital alphabetic words, small alphabetic words or words starting with capital alphabets; all these words are treated as different words. Maybe the words and their meanings are the same, but because of their writing style, they consider as distinct during processing. So we converted the whole dataset of ours into small alphabetic words to remove this confusion from our dataset. This step is also performed in R-studio.

Stop Words Removal: There is a proper list of words that are called stop words, e.g. and, the, that, which, etc. These words completed the sentences and gave them adequate meaning and make sense of the sentence. But in our research we are not processing the data as sentences; we are working on the dataset as separate words which means stop words are not part of our study, it will only make data big. So we removed all the stop words using R-studio.

Removal of Single Alphabets: After performing all the above step, we noticed that there are single letter words are present in our dataset, which are the reminders of an apostrophe. These alphabets are of no use and only make dataset noisy, so we removed all the alphabets. This step is performed in R-studio.

Removal of Extra White Spaces: We have separated the words with single space but after completing the above steps and removing all the noisy words; a lot extra white spaces have appeared in the dataset, so we removed all these additional white spaces. This last step of cleaning our dataset is also performed in R-studio.

3.3 Polarity scoring

In this step, we calculate polarity for our dataset. The polarity scores of preprocessed data are computed from Vader (Valence Aware Dictionary and sentiment Reasoner). It is an open-source python library used for sentiment analysis. The Vader dictionary performs outstandingly well in the domain of social media. As we use the lexicon approach, so Vader also belongs to SA, which is based on the words related to the sentiment lexicons. Vader not merely does necessary coordinating among the words in the content and its dictionary. It additionally considers a few things regarding the way words are composed and their context. Each word in Vader is assigned some numerical values as polarity, which may be positive or negative. We extract sentiment words (Negative or Positive) and discard words that have zero polarity by using Vader. We found almost 3000 words that have a positive or negative polarity. We also calculate polarity from another tool, i.e. TextBlob. It is also a python library used for the processing of textual data. But after evaluation, TextBlob gives wrong results for our dataset, so we don't use it for our further steps. Accuracy comparison between TextBlob and Vader is shown in fig 2.

3.4 Polarity comparison

Here comes the final and deciding step of this research. We assign polarity to our dataset as mentioned in the above section, but we must compare this dataset and its polarity with any general language dictionary which contains words alongside their polarity in general mean not specific to any domain.

We used dictionary Senticnet 4.0, lexical resource expressly conceived for supporting opinion mining and sentiment classification applications. Senticnet 4.0 is an enhanced adaptation of Senticnet 3.0, a lexical resource publicly accessible for research purposes, authorized to many research groups and utilized as a part of an assortment of research ventures around the world. It is essential to state now that we are not intended in extricating all terms, just those that demonstrate positive or negative inside the specific area, and, all the more significantly only when their introduction contrasts from the one they display when all is said in different domains or general language. Because of this point of view, terms, for example, expert, investor or doctor are unimportant to us, because these are neutral.

Our method to distinguish applicable terms is as follows:

1. Check the semantic introduction of every applicant term in our dataset by investigating them in context.
2. Removed neutral terms; means those terms whose importance does not pass on a specific semantic introduction.
3. Compare the list of the polarized words against our current list of the polarized words.
4. Removed words whose polarity is similar to our current general language words.
5. The rest of the terms are approved as domain-specific words.

3.5 Strong and weak polarity

This research gave us one bonus result. After we assigned polarity to all words accordingly to their domains and compare this polarity to general language dictionary to extract those words whose polarity complete differ from positive to negative and vice versa from our lexicon to general language dictionary; we observed that there are multiple words in our lexicon and general language dictionary which stay positive or negative in both cases, but their polarity differs. We analyzed that some words polarity in our lexicon is higher than polarity in general language dictionary, we called those words ‘strongly positive or negative’. Whereas some words polarity in our lexicon is less than polarity in general language dictionary, we called those words ‘weakly positive or negative’. This factor is present in all the domains polarity as a single word may be positive or negative in all the domains but differ in polarity because some words are not significantly crucial in one domain but may have significant importance in other domains.

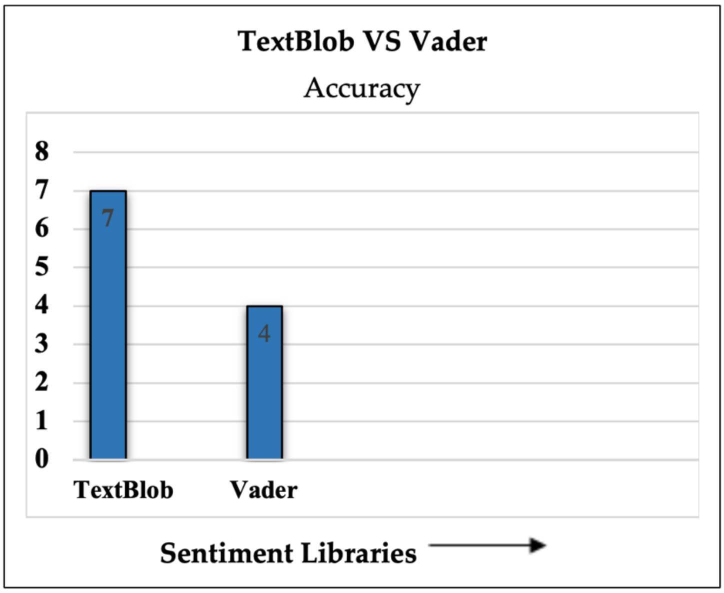

Fig. 2. Comparison between TextBlob and Vader

4 Results and Analysis

The discovering of area particular words for various areas increment our insight as well as helps us to comprehend these words in various courses of life. We characterize the polarity of our dataset concerning its domain and compare it with a general language dictionary to extract domain-specific words.

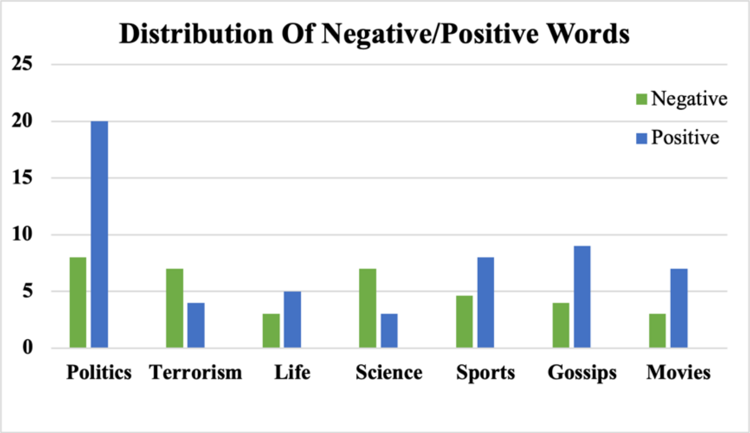

Fig. 3. Extracting semantic words as positive or negative form Vader

This procedure is not free of its issues, however. Though some lexical things unmistakably have a place in the specific talk, others, for instance, the recognized term subsidence, broadly utilized as a part of the area of a fund with a stamped semantic introduction, is likewise found in the general language dictionary, so there is no compelling reason to incorporate it in the particular vocabulary. We need to state this was the situation with numerous different terms: they were once constrained to specific vocabulary, in any case, as of late, without a doubt because of the worldwide money related circumstance, they have been progressively advancing toward the general dictionary, as they have been frequently utilized as a part of general-group of onlookers media and given uncommon consideration by the overall population, who is presently very (and unfortunately) comfortable with terms, for example, lodging air pocket or credit crunch.

Our dataset about politics contained 7744 words, 7717 words for terrorism, 7040 words for life domain, 7449 words for science, 7734 words for sports, 7091 words for gossip and 7340 words for the movie domain. Fig 4(a) and fig 4(b) depict the percentage of general and domain-specific words within the data, respectively. The data is divided into positive, negative and neutral words. The comparative analysis showed us that there are 68% words in politics domain whose polarity differs from general language words domain and same in other cases, but the point we have noted in this analysis is mostly words remained positive or negative in our domain as well as in general words domain only their polarity showed slightly changed because some words have more significance in the certain domain than general and vice versa. But we extracted those words from this work whose meaning or polarity completely different in comparison of both domains; mean if a word is positive in politics domain then in the general domain it will be negative. And we did it for all the cases of our data as shown in Table 1.

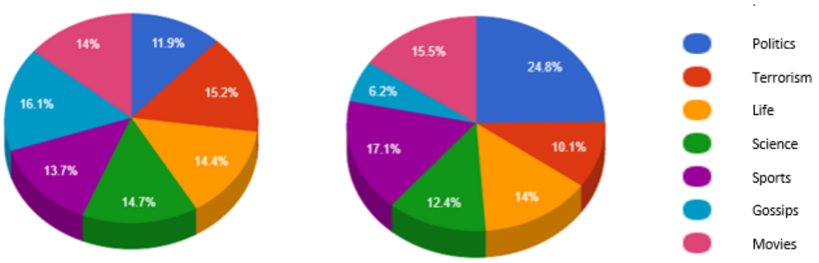

Fig. 4. Fig. 4 (a). General Words Fig. 4(b). Domain Specific Words

4.1 Summary of result

The total number of positive and negative words in a domain after assigning domain-specific polarity is separated, as shown in Fig 3. Now the words have polarity according to their specific domains, and on this basis, we extracted how much positive and negative words occurrence present in each domain. The comparative analysis is done in each domain case to analyze the difference of occurrence of domain-specific words and general words in every domain. This analysis shows that how much defining a word under its domain changes its meaning in that domain compare to its meaning in general language dictionary where all words are defined in general mean without under any circumstances, boundaries and domain.

The domain-specific words in all domains (within our dataset) which change entirely polarity from positive to negative and vice versa in comparison to general word dictionary are given in Table 2 along with their polarity status. We can see in Table 2 even the same words in different domain completely change in polarity, which changes the whole and sole meaning of the word.

Table 2. Domain Specific Words

Sr#	Politics	Terrorism	Life	Science	Sports	Movies
1	Pay (-)		Pay (+)		Pay (-)	Pay (-)
2	Growth (+)	Growth (-)	Growth (+)	Growth (-)		
3	Increase (-)	Increase (-)	Increase (+)			
4	Justice (+)	Justice (-)	Justice (+)		Justice (+)	
5	Intense (-)	Intense (-)	Intense (+)		Intense (-)	Intense (-)
6	Defense (+)	Unemployment (+)	Cutting (-)	Amazon (+)		Offend (-)
7	Leave (-)	Force (+)		Apologize (+)	Domination (-)	Obsess (-)
8	Overwhelmingly (-)		Looser (-)	Lucky (+)	Dominates (+)	
9	Killed (-)	Defense (+)	Leave (-)	Increase (+)	Certainly (+)	
10	Tough (-)	Thriller (+)	Number (+)	Excited (+)	Lucky (+)	
11			Drop (-)	Energy (+)	Exclusive (+)	
12	Lying (+)		Offend (+)	Rigorous (-)	Thrillers (+)	
13			Excuse (+)		Silly (+)	

5 Conclusion

Sentiment Analysis is defined as the process of recognizing and classifying a person’s opinions explain in the comment or message to find out whether the person’s response towards the item or product is positive, negative or neutral. A domain-specific language is defined as specification language committed to a specific issue in the domain and a particular description of issue method or a particular issue results. The purpose of this research was to identify words that are interpreted differently in various fields. Polarity had been given to words of the dataset according to their domains and based on this polarity. They have been recognized as positive negative and neutral and evaluated as domain-specific words. Multiple libraries were also used in the scoring process of the dataset. This research improved the knowledge about the context-dependency and found the core-specific meanings of words in multiple fields and interprets those words based on their domain-specific meanings. It will enhance the performance of evaluating sentiment analysis for these domains.

In the future, instead of identifying the words for domain-specific, we will work on identifying the domain of dataset first. We will increase the data set and apply different algorithms for scoring and try to use other techniques for sentiment analysis.

Also, new and better approaches to data mining will be applied for the analysis of results.

6 Acknowledgement

We thank Library of Norwegian University of Science and Technology, Norway for paying the Open Access Charges to publish this paper.

7 References

[1] Vinodhini, G., & Chandrasekaran, R. M. "Sentiment analysis and opinion mining: a survey." International Journal 2, no. 6 (2012): 282-292.

[2] Schouten, K., & Frasincar, F. "Survey on aspect-level sentiment analysis." IEEE Transactions on Knowledge and Data Engineering 28, no. 3 (2016): 813-830. https://doi.org/10.1109/tkde.2015.2485209

[3] Jiao, J., & Zhou, Y. "Sentiment polarity analysis based multi-dictionary." Physics Procedia 22 (2011): 590-596. https://doi.org/10.1016/j.phpro.2011.11.091

[4] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Urena-Lopez, L. A. "Random walk weighting over sentiwordnet for sentiment polarity detection on twitter." In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 3-10. Association for Computational Linguistics, 2012. https://doi.org/10.1016/j.csl.2013.04.001

[5] Turney, P. D. "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. https://doi.org/10.3115/1073083.1073153

[6] Kim, S. M., & Hovy, E. "Determining the sentiment of opinions." In Proceedings of the 20th international conference on Computational Linguistics, p. 1367. Association for Computational Linguistics, 2004. https://doi.org/10.3115/1220355.1220555

[7] Hiroshi, K., Tetsuya, N., & Hideo, W. "Deeper sentiment analysis using machine translation technology." In Proceedings of the 20th international conference on Computational Linguistics, p. 494. Association for Computational Linguistics, 2004. https://doi.org/10.3115/1220355.1220426

[8] Kennedy, A., & Inkpen, D. "Sentiment classification of movie reviews using contextual valence shifters." Computational intelligence 22, no. 2 (2006): 110-125. https://doi.org/10.1111/j.1467-8640.2006.00277.x

[9] Devitt, A., & Ahmad, K. "Sentiment polarity identification in financial news: A cohesion-based approach." In ACL, vol. 7, pp. 1-8. 2007.

[10] McDonald, R., Hannan, K., Neylon, T., Wells, M., & Reynar, J. "Structured models for fine-to-coarse sentiment analysis." In Annual meeting-association for computational linguistics, vol. 45, no. 1, p. 432. 2007.

[11] Blitzer, J., Dredze, M., & Pereira, F. "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification." In ACL, vol. 7, pp. 440-447. 2007.

[12] Chaumartin, F. R. "UPAR7: A knowledge-based system for headline sentiment tagging." In Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 422-425. Association for Computational Linguistics, 2007. https://doi.org/10.3115/1621474.1621568

[13] Baccianella, S., Esuli, A., & Sebastiani, F. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, pp. 2200-2204. 2010. https://doi.org/10.7717/peerj-cs.252/fig-9

[14] Kando, N. "Overview of the Fifth NTCIR Workshop." In NTCIR. 2005.

[15] Kando, N. "Overview of the Seventh NTCIR Workshop." In NTCIR. 2008.

[16] Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. "Twitter sentiment analysis for large-scale data: an unsupervised approach." Cognitive computation 7, no. 2 (2015): 254-262. https://doi.org/10.1007/s12559-014-9310-z

[17] Cambria, E., Havasi, C., & Hussain, A. "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis." In FLAIRS conference, pp. 202-207. 2012.

[18] Cambria, E., Olsher, D., & Rajagopal, D. "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis." In Twenty-eighth AAAI conference on artificial intelligence. 2014.

[19] Cambria, E., & Hussain, A. Sentic computing: Techniques, tools, and applications. Vol. 2. Springer Science & Business Media, 2012.

[20] Cambria, E., Benson, T., Eckl, C., & Hussain, A. "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality." Expert Systems with Applications 39, no. 12 (2012): 10533-10543. https://doi.org/10.1016/j.eswa.2012.02.120

[21] Pak, A., & Paroubek, P. "Twitter as a corpus for sentiment analysis and opinion mining." In LREc, vol. 10, no. 2010. 2010.

[22] Zhou, X., Tao, X., Yong, J., & Yang, Z. "Sentiment analysis on tweets for social events." In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, pp. 557-562. IEEE, 2013. https://doi.org/10.1109/cscwd.2013.6581022

[23] Milstein, S., Lorica, B., Magoulas, R., Hochmuth, G., Chowdhury, A., & O'Reilly, T. Twitter and the micro-messaging revolution: Communication, connections, and immediacy--140 characters at a time. O'Reilly Media, Incorporated, 2008.

[24] Asur, S., & Huberman, B. A. "Predicting the future with social media." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 492-499. IEEE, 2010. https://doi.org/10.1109/wi-iat.2010.63

[25] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. "Predicting elections with twitter: What 140 characters reveal about political sentiment." Icwsm 10, no. 1 (2010): 178-185. https://doi.org/10.1177/0894439311404119

[26] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. "Lexicon-based methods for sentiment analysis." Computational linguistics 37, no. 2 (2011): 267-307. https://doi.org/10.1162/coli_a_00049

[27] Mostafa, M. M. "More than words: Social networks' text mining for consumer brand sentiments." Expert Systems with Applications 40, no. 10 (2013): 4241-4251. https://doi.org/10.1016/j.eswa.2013.01.019

[28] Fan, M., & Wu, G. "Aspect opinion mining on customer reviews." ICCE2011. AISC 112 (2011): 27-33.

[29] Saif, H., He, Y., & Alani, H. "Semantic sentiment analysis of twitter." The Semantic Web–ISWC 2012 (2012): 508-524. https://doi.org/10.1007/978-3-642-35176-1_32

[30] Kumar, A., & Teeja, M. S. "Sentiment analysis: A perspective on its past, present and future." International Journal of Intelligent Systems and Applications 4, no. 10 (2012): 1.

[31] Pang, B., Lee, L., & Vaithyanathan, S. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical

methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002. https://doi.org/10.3115/1118693.1118704

[32] Kaushik, C., & Mishra, A. "A scalable, lexicon based technique for sentiment analysis." arXiv preprint arXiv:1410.2265 (2014).

[33] Soni, S., & Sharaff, A. "Sentiment Analysis of Customer Reviews Based on Hidden Markov Model." In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), p. 12. ACM, 2015. https://doi.org/10.1145/2743065.2743077

[34] Lu, T. J. "Semi-supervised microblog sentiment analysis using social relation and text similarity." In Big Data and Smart Computing (BigComp), 2015 International Conference on, pp. 194-201. IEEE, 2015. https://doi.org/10.1109/35021bigcomp.2015.7072831

[35] Arora, D., Li, K. F., & Neville, S. W. "Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study." In Advanced Information Networking and Applications (AINA), 2015 IEEE 29th International Conference on, pp. 680-686. IEEE, 2015. https://doi.org/10.1109/aina.2015.253

[36] Do, H. J., & Choi, H. J. "Sentiment analysis of real-life situations using location, people and time as contextual features." In Big Data and Smart Computing (BigComp), 2015 International Conference on, pp. 39-42. IEEE, 2015. https://doi.org/10.1109/35021bigcomp.2015.7072847

[37] Pawar, Kishori K., Pukhraj P. Shrishrimal, and R. R. Deshmukh. "Twitter sentiment analysis: A review." International Journal of Scientific & Engineering Research 6.4 (2015): 9.

[38] Medhat, W., Hassan, A., & Korashy, H. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5, no. 4 (2014): 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

[39] Guzman, E., & Maalej, W. "How do users like this feature? a fine grained sentiment analysis of app reviews." In Requirements Engineering Conference (RE), 2014 IEEE 22nd International, pp. 153-162. IEEE, 2014. https://doi.org/10.1109/re.2014.6912257

[40] Bansal, M. "SENTIMENT ANALYSIS FROM SOCIAL MEDIA LIVE FEEDS USING UNSTRUCTURED DATA MINING."

[41] Park, S., Lee, W., & Moon, I. C. "Efficient extraction of domain specific sentiment lexicon with active learning." Pattern Recognition Letters 56 (2015): 38-44. https://doi.org/10.1016/j.patrec.2015.01.004

[42] Khuc, V. N. "Approaches to Automatically Constructing Polarity Lexicons for Sentiment Analysis on Social Networks." PhD diss., The Ohio State University, 2012.

[43] Agarwal, B., Mittal, N., Bansal, P., & Garg, S. "Sentiment analysis using common-sense and context information." Computational intelligence and neuroscience 2015 (2015): 30. https://doi.org/10.1155/2015/715730

[44] Trung, D. N., Jung, J. J., & Kiss, A. "Towards modeling fuzzy propagation for sentiment analysis in online social networks: A case study on TweetScope." In Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, pp. 331-338. IEEE, 2013. https://doi.org/10.1109/coginfocom.2013.6719266

[45] Hussain, A., Sattar, S., & Afzal, M. T. "Literature Review on Feature Identification in Sentiment Analysis." International Journal of Computer Applications 132, no. 3 (2015): 22-27. https://doi.org/10.5120/ijca2015907331

[46] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. "Ontology-based sentiment analysis of twitter posts." Expert systems with applications 40, no. 10 (2013): 4065-4074. https://doi.org/10.1016/j.eswa.2013.01.001

[47] Choi, Y., Kim, Y., & Myaeng, S. H. "Domain-specific sentiment analysis using contextual feature generation." In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 37-44. ACM, 2009. https://doi.org/10.1145/1651461.1651469

[48] Koto, F., & Adriani, M. "The use of POS sequence for analyzing sentence pattern in Twitter sentiment analysis." In Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on, pp. 547-551. IEEE, 2015. https://doi.org/10.1109/waina.2015.58

[49] Team, R. D. C. R: a language and environment for statistical computing R Foundation for Statistical Computing, 2.13. Vienna, Austria, 2011. ISBN 3-900051-07-0, URL http://www. R-project. org.

8 Authors

Kamran Shaukat is a Ph.D. student at The University of Newcastle, Australia. kamran.shaukat@uon.edu.au

Ibrahim A. Hameed is a full professor at Norwegian University of Science and Technology, Norway.

Suhuai Luo is an Associate Professor at The University of Newcastle, Australia.

Imran Javed is an Assistant Professor at University of the Punjab, Pakistan.

Farhat Iqbal, Amber Faisal, Rabia Masood, Ayesha Usman, Usman Shaukat, Rosheen Hassan, Aliya Younas, Shamsair Ali, Ghazif Adeem are students at University of the Punjab, Pakistan.

Article submitted 2020-01-09. Resubmitted 2020-02-18. Final acceptance 2020-02-19. Final version published as submitted by the authors.