ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 10 (2020) No.3, pp. 140-150
DOI:10.33168/JSMS.2020.0310

A Combination of Lexicon and Machine Learning Approaches for Sentiment Analysis on Facebook

Alaa Thamer Mahmood[1], Siti Sakira Kamaruddin[2], Raed Kamil Naser[3],
Maslinda Mohd Nadzir[4]

[1] Technical Instructors Training Institute, Middle Technical University, Baghdad, Iraq
[2,4] School of Computing, Universiti Utara Malaysia, Kedah, Malaysia
[3] Administration Directorate, Ministry of Defense organization, Baghdad, Iraq

Abstract. The increase of user-generated content (UGC) on the Internet has led previous studies to propose various sentiment analysis approaches to understand public opinion. The primary goal is to enhance engagement through social media by analyzing various feedback. Sentiment analysis is performed based on two approaches i.e. machine learning and lexicon-based. Since approaches based on machine learning require costly preparation of training dataset and the approaches based on lexicon produce unsatisfactory performance, in this paper, both approaches are combined to perform sentiment analysis on Facebook comments. The importance of a lexicon-based approach to automatically construct the labeled data for machine learning sentiment classification is discussed in this paper. Experiments performed using the Universiti Utara Malaysia (UUM) Facebook posts show that using the combined lexicon-based and machine learning approach on two classifiers i.e. Naïve Bayes and Support Vector Machine outperform the single approaches to produce more accurate sentiment classifications.

Keywords: Machine learning, lexicon, sentiment analysis, social media, Facebook, text mining.

1. Introduction

The volume of textual documents in social media data has promoted the growth of user-generated content (UGC) that is widely used among decision-makers to provide cues of information in a certain context (Jovanoski, Pachovski, and Nakov 2015). Also, the growth of UGC on the Internet has led previous studies to propose various sentiment analysis approaches (Lei & Xin, 2011, Yang, 2013, Subramaniyaswamy et al. 2017, Al-Ayyoub, 2019, Habimana et al. 2020 and Ruz, Henríquez, and Mascareño 2020).

Social media-based sentiment analysis is appealing to researchers and has been utilized for various applications (Habimana et al. 2020, Subramaniyaswamy et al. 2017 and Ruz, Henríquez, and Mascareño 2020) however, there is limited work that performs the comparison between the major approaches that are available in the sentiment analysis field. Thus, it is evident from the literature such a comparison is needed to provide a guideline for future researchers in the field.

Sentiment analysis deals with subjective information which refers to users' points of view, such as feelings and perceptions about certain events (Ruz, Henríquez, and Mascareño 2020). Analyzing sentiments in social media sites would help to convey the attitudes of users or members of the public towards a certain situation. Hence, increasing the organizational control in identifying new opportunities and determine the settings to suit users' expectations (Chenghua Lin 2014).

The major challenge of analyzing sentiments is in terms of the complexity of natural language text. This is the reason why researchers working on text prefer to use complex graph-based approaches to analyze and represent text as shown in Kamaruddin et al. (2009) and Abdulsahib and Kamaruddin, (2015). The difficulties to track and measure the unorganized and huge size of data accessible in most social networking sites such as Facebook often hinder work in this area (Prichard et al. 2015). Monthly active users of Facebook are approximated to be 1.59 billion (Hassani et al., 2020). Facebook has been used in educational settings as depicted in various studies (Teck et al. 2013; Zamani et al. 2014; Ortigosa, Martín, and Carro 2014; Saykili and Kumtepe 2019 and Barrot 2018).

Performing sentiment analysis in Facebook needs quantitative summaries for posts, such as general capacity and valence associated with the average of rating content for users, to represent the users' opinions. Work such as in (Zamani et al. 2014)] considered the potential of Facebook to provide an overview of people's opinions by mining and performing sentiment analysis on Facebook comments. Universiti Utara Malaysia (UUM) has been using Facebook as an effective way to allow students to express their opinion and comments about different services (Teck et al. 2013), however, the information was not harnessed to enable the university management in making strategic decisions.

Concerning sentiment analysis approaches, two primary approaches exist i.e. the machine learning and the lexicon-based approach. The machine learning-based approach uses manually labeled data to train the classifiers (Ruz, Henríquez, and Mascareño 2020). The learning algorithms rely on the coverage and quality of the training data, hence it is more labor-intensive and higher cost as opposed to the lexicon-based approach. However, in terms of performance machine learning approach usually outperforms the lexicon-based approach.

The lexicon-based approach analyzes the text and gives a sentiment score to it, by referring to a sentiment lexicon which is pre-defined (Taboada et al. 2011). One example of sentiment lexicon is the SentiWordNet (Esuli and Sebastiani 2006), nevertheless, most researchers prefer to employ their own lexicon which is typically constructed based on the analyzed text. A lexical database, for example, WordNet (Miller 1995) are often used for this purpose. The superiority of the approach based on lexicon is, it does not necessitate the requirement to prepare a labeled data that are used for training in the approaches which are based on machine learning.

This paper describes a sentiment analysis effort for Facebook comments using a combination of lexical and machine learning-based approaches. In section 2, we present the related work on sentiment analysis. Section 3 describes the experimental design of this work. While section 4 discusses the analysis of the results. Finally, in section 5, we present the conclusions.

2. Related Works

The massive use of social media sites made it one of the main sources for understanding people's behavior stimulated when people are communicating and sharing thoughts or opinions online (Neri et al. 2012).

For example, (Pang and Lee 2008) surveyed different sentiment analysis approaches for promoting the applications that leverage the potentials of identifying sentiments, as compared to other applications that are more fact-based. They concluded that the polarity-based datasets should always consider the middling ratings in practice and acknowledged the potential of features based on linguistics to analyze the Twitter sentiments (Kouloumpis, Wilson, and Moore 2011). This approach was typically used to gather the necessary information about the informal and creative language in microblogging. The work in (Khan, Atique, and Thakare 2015) proposed an entity-level sentiment analysis method for Twitter by utilizing a lexical-based approach.

The work in (Troussas et al. 2013) applied sentiment analysis for assessing language learning in the Facebook platform. The work in (Trinh et al. 2016) proposed the development of content-based ranking by considering the association between the comment polarities in which users' comments on Facebook are all analyzed using a lexicon-based approach based on the social packets crawler. (Habernal, Ptáček, and Steinberger 2014) asserted the need for developing machine learning approaches for sentiment analysis on the Facebook platform. They reasoned that such limitation is due to the lack of entity recognition and pre-processing on sentiment classification.

Owing to the increase of information available in social media, researchers are concentrating on implementing sentiment analysis to this information as discussed above, however, most of these works retrieve text from sources such as Twitter. Obtaining data from Twitter requires less effort compared to retrieving from other social network sources such as Facebook.

Sentiment analysis is performed based on two principle approaches i.e. machine learning and lexicon. The usage of machine learning-based approach for sentiment analysis on user reviews of movies is portrayed in Pang, Lee, and Vaithyanathan (2002). They carried out a comparison of different classifiers to classify sentiments in movie reviews. They recorded 82.9% accuracy using Support Vector Machines (SVM). In specific domains, such as movie review, the machine-learning-based approach performs well.

In Taboada et al. (2011) an approach based on lexicon was used to perform sentiment analysis on six textual datasets from various domains. They recorded 75–80% accuracy. As mentioned in the previous section approaches based on lexicon has an advantage because it does not depend on a labeled training set to classify the sentiments. However, the lexicon-based approach usually does not perform well in specific domains compared to text that is not bounded by domains. The reason for this is the lexicon-based approach uses a rule-based method to check against pre-defined lexicons whereas a machine learning-based method learns the patterns from the given training data. If the training data is comprehensive and covers all possible variety of data features the machine learning algorithm will yield more accurate results.

From the above mentioned related works, it is possible to deduce that the machine learning approach obtains better results in the bounded domains. However, to be applied to the general domain, it requires training datasets that are pre-labeled, thus features identifying the neutral, positive, and negative polarity are learned. In contrast, a lexicon-based approach needs a polarity dictionary that contains words and phrases which are annotated by its semantic orientation. However, dictionaries are extensible and more available than training datasets. These approaches, when applied on bounded domains, are less accurate than the machine-learning approaches, but when applied across all domain, it is more robust.

Therefore, we propose to combine both the lexicon and machine learning-based approaches to perform sentiment analysis on UUM Facebook comments. However, the combination was done differently compared to existing works such as Khan, Atique and Thakare (2015) which are explained in detail in the next section.

3. Experimental Design

This section discusses the execution of the proposed approach and the steps taken to build the combined lexicon and machine learning-based approaches. Specifically, the combined approaches are supervised machine learning and a dictionary-based approach.

Fig. 1 illustrates the mains steps of the experimental setting for the combined machine learning and lexical-based approaches for analyzing the sentiments from UUM Facebook comments.

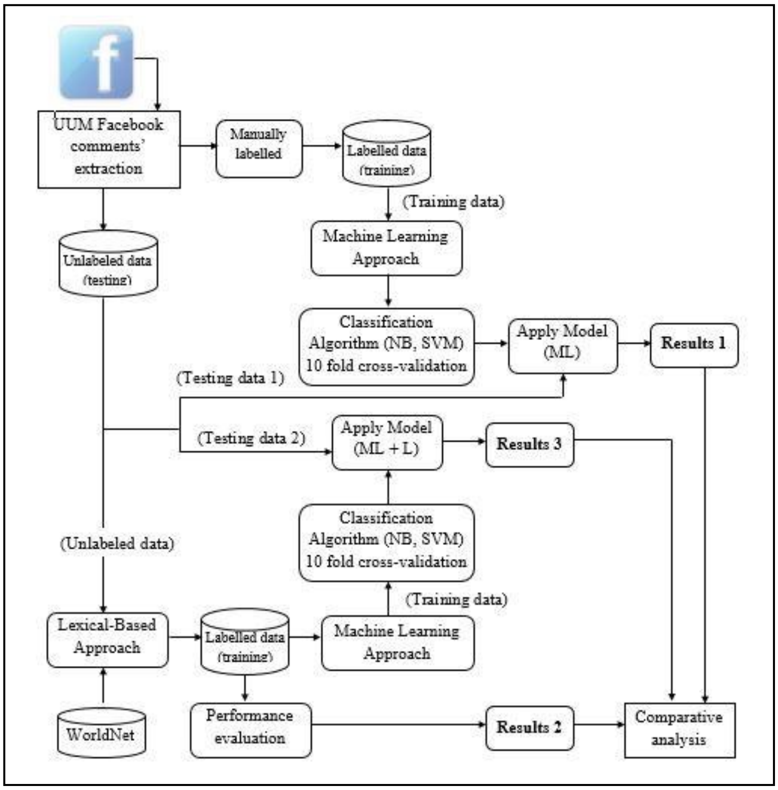

Fig. 1.Combined lexicon and machine learning-based approach for sentiment analysis.

The experiment starts with the first step of extracting UUM Facebook comments. The comments were extracted via the Facepager (Surroop, Canoo, and Pudaruth 2016) tool. Facepager is a scraper tool to extract data that is openly available on Twitter and Facebook. All the obtained data are kept in a CSV file. The dataset contains the UUM Facebook page comments (unlabeled data).

For the machine learning approach, the data were divided into two i.e. the training data and the testing data. The training data were labeled manually as having a positive and negative polarity (labeled data). Next, the labeled data were given as input to the machine learning approach by training the two classifiers; Naïve Bayes and Support Vector Machine (SVM). Naïve Bayes has been successfully used for sentiment classification as shown in Troussas et al., (2013) and Alkubaisi, Kamaruddin and Husni (2018), whereas SVM is often used in sentiment analysis as shown in Pang, Bo, and Lillian Lee. (2008) and Kaswidjanti, Himawan, and Silitonga (2020). SVM works by isolating the search space by the greatest margin hyperplane that results in the dividing of the training data into two distinguishable classes (Yusof, Mohamed, and Abdul-Rahman,2015; Yusof et al., 2015). Then the produced model was tested using the testing data (unlabeled data). The produced result is identified as Results 1.

For the lexicon-based approach, the unlabeled data were entered into the nominal to text operation to alter the type of nominal attributes. After this operation, the unlabeled data were pre-processed which includes operations such as tokenization, transform cases, and filter tokens to clean the text from non-useful data. After the completion of the pre-processing step, the sentiment was extracted. This extraction was to determine the sentiment for each unlabeled word, done by checking the word against the Wordnet dictionary. Finally, the output is the labeled data, with positive or negative polarity. The produced result is identified as Results 2.

In the combined lexicon-based and machine learning approach, the output from the lexicon-based approach was used in the machine learning approach as a training data to train the two classifiers (Naïve Bayes, SVM classifiers). Then, the same testing data (unlabeled data) was used to examine the performance of the two classifiers which provided Results 3.

The evaluation of the study was based on the measurement of the performance accuracy (1) of the proposed combination along with the precision value (2) and recall (3) and F-measure (4) where TP represents true positive, TN represents the true negative, FP represents false positive and FN represents false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F - measure = \frac{2 \times Precision \times Recall}{2 \times Precision + Recall} \tag{4}$$

Finally, all the produced results (Result 1, Results 2, and Results 3) were analyzed and compared.

4. Results and Discussion

This section presents the results achieved via experiments conducted on UUM Facebook comments. The results are presented in Table 1.

TABLE I. RESULTS OF THE THREE APPROACHES

Approach	**Classifier**	**Performance Measure (percentage)**			
		Accuracy	***Recall***	***Precision***	***F-Measure***
Machine learning	Naïve Bayes	73	73	91	81
	SVM	80	100	79	88
Lexicon Based	-	85	91	88	91
Combined Lexicon and ML	Naïve Bayes	86	87	96	91
	SVM	**90**	**100**	**90**	**94**

The table shows the results for the accuracy of the three approaches: Machine Learning, Lexicon-Based, and Combined Lexicon and Machine Learning Approaches.

The lowest accuracy percentage is 73%, represented by the machine learning approach using the Naïve-Bayes algorithm before combining with the lexicon-based approach. It reached a higher percentage of 86% after combining the lexicon and machine learning approaches. The result of the lexicon-based approach alone was 85%.

These results indicate that a combined machine learning and lexicon-based approach is a better approach compared to Machine Learning Approach using training data that was labeled manually. Hence, the lexicon-based approach can be utilized to label data automatically and obtain meaningful training data, and thus increasing classification performance for sentiment analysis.

5. Conclusion

The performance of sentiment analysis partly depends on the data set if there is a clearly labeled positive and negative polarity identification. The lexicon-based approach does not need training data since it depends on dictionaries; otherwise, the machine learning approach works very well if the training data has clear labeled positive and negative polarity. As shown in this paper, the benefits of both approaches can be leveraged by combining them, more specifically, the data were automatically labeled using a lexicon-based approach before using it as the training data in the machine learning approach. This can reduce substantially the cost of preparing complete training data for use in the machine learning approach.

6. Acknowledgment

This publication is sponsored by the Ministry of Education Malaysia (Fundamental Research Grant Scheme, FRGS-13259).

References

Abdulsahib, A. K., & Kamaruddin, S. S. (2015). “Graph based text representation for document clustering”. *Journal of Theoretical and Applied Information Technology*, *76*(1), 1-13.

Alkubaisi, G.A.A.J., Kamaruddin, S.S. and Husni, H., (2018). “Conceptual framework for stock market classification model using sentiment analysis on twitter based on Hybrid Naïve Bayes Classifiers.” *International Journal of Engineering & Technology*, *7*(2.14), pp.57-61.

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). “A comprehensive survey of arabic sentiment analysis”. *Information processing & management*, *56*(2), 320-342.

Barrot, J. S. (2018). “Facebook as a Learning Environment for Language Teaching and Learning: A Critical Analysis of the Literature from 2010 to 2017.” *Journal of Computer Assisted Learning* 34(6): 863–75.

Chenghua L., Yulan H. (2014). “Sentiment Analysis in Social Media.” *Encyclopedia of Social Network Analysis and Mining*: pp 1688-1699.

Esuli, A., and Sebastiani, F. (2006). “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.” *Proceedings of the 5th Conference on Language Resources and Evaluation*: pp. 417–22. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217.

Habernal, I, Ptáček T., and Steinberger, J. (2014). “Supervised Sentiment Analysis in Czech Social Media.’” *Information Processing and Management* 50(5): 693–707.

Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2020). "Sentiment analysis using deep learning approaches: an overview". *Science China Information Sciences*, 63(1), 1-36.

Hassani, H., Beneki, C., Unger, S., Mazinani, M.T. and Yeganegi, M.R., (2020). "Text Mining in Big Data Analytics." *Big Data and Cognitive Computing*, *4*(1), p.1.

Jovanoski, D., Pachovski V., and Nakov P.. (2015). “Sentiment Analysis in Twitter for Macedonian.” *International Conference Recent Advances in Natural Language Processing, RANLP* 2015–Janua: pp. 249–57.

Kamaruddin, S.S., Hamdan, A.R., Bakar, A.A. and Nor, F.M., (2009). "Conceptual graph interchange format for mining financial statements". In *International Conference on Rough Sets and Knowledge Technology* pp. 579-586. Springer, Berlin, Heidelberg.

Kaswidjanti, W., Himawan, H., & Silitonga, P. D. P. (2020). "The accuracy comparison of social media sentiment analysis using lexicon based and support vector machine on souvenir recommendations." *Test Engineering and Management, 82*(3-4), 3953-3961.

Khan, A. Z. H., Atique M., and Thakare, V. M. (2015). “Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis.” *Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering (IJECSCSE)*: 89–96.

Kouloumpis, E, Wilson, T, and Moore, J. (2011). “Twitter Sentiment Analysis: The Good the Bad and the Omg!” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, USA.*: 538–41. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251 ?iframe=true& width=90%25&height=90%25.

Lei, G., & Xin, G. (2011). "Social network analysis on knowledge sharing of scientific groups." *Journal of System and Management Sciences*, *1*(3), 79-89.

Miller, George a. (1995). “WordNet: A Lexical Database for English.” *Communications of the ACM* 38(11): 39–41.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). “Sentiment Analysis on Social Media.” *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*: 919–26.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). “Sentiment Analysis in Facebook and Its Application to E-Learning.” *Computers in Human Behavior* 31(1): 527–41.

Pang, B., and Lee L. (2008). “Opinion Mining and Sentiment Analysis.” *Foundations and Trends in Information Retrieval* Vol. 2, No: 1–135.

Pang, Bo, Lee, L., and Vaithyanathan, S. (2002). “Thumbs up?: Sentiment Classification Using Machine Learning Techniques.” *Empirical Methods in Natural Language Processing (EMNLP)* 10(July): 79–86.

Prichard, J., Watters, P., Krone, T., Spiranovic, C., & Cockburn, H. (2015). “Social

Media Sentiment Analysis: A New Empirical Tool for Assessing Public Opinion on Crime?” *Current Issues in Criminal Justice* 27(2): 217–26.

Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). “Sentiment Analysis of Twitter Data during Critical Events through Bayesian Networks Classifiers.” *Future Generation Computer Systems* 106(May): 92–104.

Saykili, A., & Kumtepe, E. G. (2019). “Educational Use of Facebook: A Comparison of Worldwide Examples and Turkish Context. *In Digital Turn in Schools—Research, Policy,* Practice (pp. 251-267). Springer, Singapore.

Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., & Umamakeswari, A (2017). “Sentiment Analysis of Tweets for Estimating Criticality and Security of Events.” *In Improving the Safety and Efficiency of Emergency Services: Emerging Tools and Technologies for First Responders* (pp. 293-319). IGI Global.

Surroop, K., Canoo, K., & Pudaruth, S. (2016). “A Novel Position-Based Sentiment Classification Algorithm for Facebook Comments.” *International Journal of Advanced Computer Science and Applications* 7(10): 261–68.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). “Lexicon-Based Methods for Sentiment Analysis.” *Computational Linguistics* 37(2): 267–307.

Teck, T. B., Michael, E., Chuen, E. M., & Keat, C. C. (2013). “The Comparison between Public and Private University Chinese Students towards Utilizing Facebook as Public Sphere for Political Discussion : A Case Study on University Utara Malaysia (UUM) and University Tunku Abdul Rahman (UTAR).” 3(8): 858–67.

Trinh, S., Nguyen, L., Vo, M., & Do, P. (2016). “Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language.” *Recent Developments in Intelligent Information and Database Systems* 642: (pp. 263-276): Springer.

Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013). "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning." In *4th International Conference on Information, Intelligence, Systems and Applications (IISA)*: 198–205.*IISA 2013*.

Yang J., "Business model analysis for Chinese social networking website". Journal of System and Management Sciences, vol. 3, no. 3, (2013), pp. 78-84.

Yusof, N. N., Mohamed, A., & Abdul-Rahman, S. (2015). “Reviewing classification approaches in sentiment analysis” *International Conference on Soft Computing in Data Science, SCDS 2015*: (pp. 43-53). Springer, Singapore.

Yusof, Y., Ahmad, F.K., Kamaruddin, S.S., Omar, M.H. and Mohamed, A.J., (2015), September. Short term traffic forecasting based on hybrid of firefly algorithm and least squares support vector machine. *In International Conference on Soft Computing in Data Science* pp. 164-173. Springer, Singapore.

Zamani, N. A. M., Abidin, S. Z., Omar, N. and Abiden, M. Z. Z. (2014). "Sentiment Analysis : Determining People's Emotions in Facebook." *Applied Computational Science* ISBN: 978-: 111–16.