Journal of Engineering and Applied Sciences 12 (3): 462-467, 2017
ISSN: 1816-949X

Review on Sentiment Analysis Approaches for Social Media Data

[1]Nur Atiqah Sia Abdullah, [1]Nurul Iman Shaari and [2]Abd Rasid Abd Rahman
[1]Faculty of Computer and Mathematical Sciences,
[2]Faculty of Communication and Media Studies, University Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia

Abstract: This study reviews sentiment analysis approaches specifically used for political research and social media data. The comparison is based on classifier, social media type, algorithm, data review and polarity classes. In this study, systematic literature review is used to explore the sentiment analysis approaches used in classifying social media data. The approaches include supervised machine learning, unsupervised learning, lexicon-based and hybridization approaches. The reviewed literatures involve data from social media such as Twitter, Email, Youtube and websites. All the approaches are evaluated and compared based on classifier, type of social media, data review and polarity classes. Based on the comparison, most of the researches use hybrid approach to classify the social media data. The algorithms in hybrid approaches are combination of lexicon-based and supervised machine learning. Most of the social media data used in these researches are extracted from Twitter. Lexicon of dictionary based and support vector machine are used for classifying political related tweets. There are also literatures involve Malay posts in social media. The past research uses social media, blog and Facebook as data. Then the sentiment analysis approaches are based on support vector machine and lexicon-based. The polarity classes involve only positive and negative or happy, unhappy and emotionless. As a conclusion, the hybrid approach of lexicon dictionary based and support vector machine is the best hybridization approach to classify the sentiment for the Malay political tweets.

Key words: Sentiment analysis, supervised learning, lexicon-based, politic, social media, Twitter

INTRODUCTION

Social media data is very important piece of data in big data analytics since most of the users are now actively connected to social media platform. The challenges of big data analytics include mining the data and getting useful insights. Most of the researches use tweets from Twitter as collection of data because of it is public and can be extracted quite easily. These tweets are sufficient in the analytics to recognize the popular topic and public opinions. Sentiment analysis is normally used to identify the opinion from someone about a topic (Glavas *et al*., 2012). It can be done by analyzing the subjectivity in text form of computational linguistics study. Basically, sentiment can be classified into polarity such as positive, negative and neutral according to its domain. Some areas use this analysis to increase the achievement by acquires, evaluate, observe and summarize of data from social media as Twitter (Makaram *et al*., 2015).

Many types of sentiment analysis approaches are available such as Support Vector Machine (SVM) and dictionary-based which are categorised into machine learning and lexicon-based approaches as shown in (Fig 1).

Supervised learning: Supervised learning is one of the machine learning approaches. It is divided into 4 main types of classifiers such as decision tree, linear, rule-based and probabilistic.

Decision tree classifier: Decision tree classifier is normally used to prepare a space of training data for hierarchical decomposition (Medhat *et al*., 2014; Nirmala *et al*., 2013). It splits the data depends on attribute value of data that represent existence words (Medhat *et al*., 2014). In classification purpose, it divides the data space recursively until the leaf nodes remaining with minimum numbers of records. There are split types of attribute values as single attribute split, similarity-based multi attribute split and dimensional-based multi attribute split (Nirmala *et al*., 2013; Pandhe and Pawar, 2015; Kaur and Saini, 2014). There is a research on monitoring

Corresponding Author: Nur Atiqah Sia Abdullah, Faculty of Computer and Mathematical Sciences,
University Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

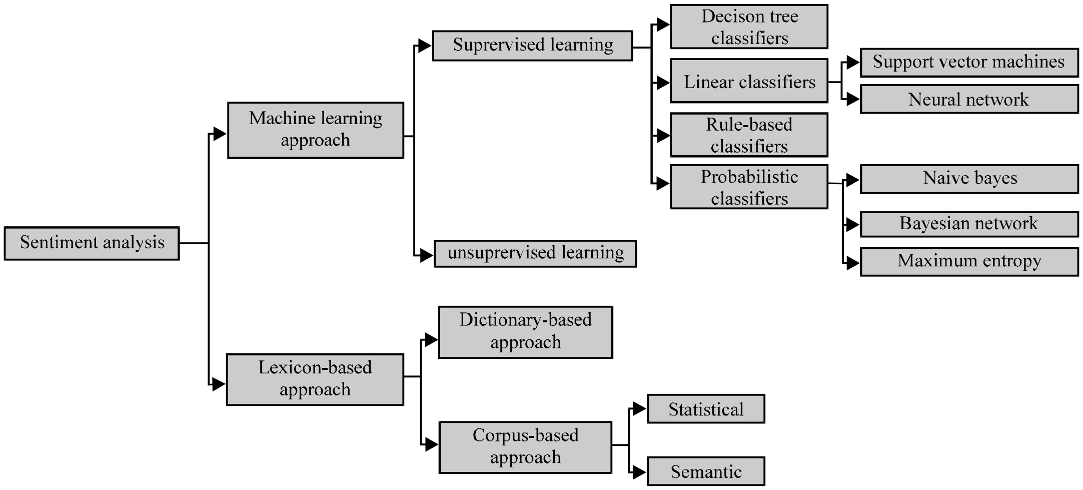

Fig. 1: Sentiment classification approaches from technical view (Medhat *et al*., 2014)

the suspicious activity related to provocative issues from social media posts (Pandhe and Pawar, 2015). The posts are classified into abusive and non-abusive categories. Iterative Dichotomiser 3 (ID3) algorithm is used in this research (Pandhe and Pawar, 2015). Additionally, ID3 is also used in previous researches on Spanish poetry formal writing and English email informal writing data (Kaur and Saini, 2014).

Linear classifier: Linear classifier can be basically divided into 2 main classifiers, namely Support Vector Machine (SVM) and Neural Network (NN) (Medhat *et al*., 2014; Nirmala *et al*., 2013; Brynielsson *et al*., 2014; Alsaffar and Omar, 2014; Patel and Mistry, 2015; Dickinson, 2015). SVM is used to determine good linear separators between different classes (Nirmala *et al*., 2013). It suits for big size of unlabeled and small size of labelled data apart of simple and interpretability. Besides, there is a research used SVM to classify the user emotion based on social media posts (Brynielsson *et al*., 2014). SVM is also used in classifying the Malay online posts data from blog and social media (Alsaffar and Omar, 2014). As a result, it found the highest performance on F1-measure achieved when using SVM compared to other approaches.

There are seven formal and 4 informal writing styles researches used SVM approach (Kaur and Saini, 2014). Based on the research, SVM gained the highest accuracy performance for formal text as shown in (Fig. 2). It proved that SVM has the most stable performance for classifying sentiment value in any language rather than other approaches (Fig. 2). Neural Network (NN) consists of many neurons where the neuron is the basic unit (Medhat *et al*., 2014). It is a self adaptive approach that able to adjust the data weight without specification.

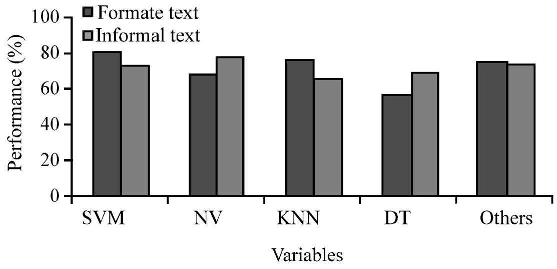

Fig. 2: Comparative performance in formal and informal of previous research (Kaur and Saini, 2014)

Two types of NN include feed forward multilayer networks and multilayer preceptrons (Nirmala *et al*., 2013). There is investigation on logical management domain based on the reviews from social websites which prove NN is applicable as classifier (Patel and Mistry, 2015). In other research, 2 classifiers are combined in detecting illegal drug usage on social media where SVM used for identify textually expressed intent to use or distribute illegal drugs on Twitter and deep neural network to recognize object in large variety.

Rule-based classifier: In this classifier, the data space is modelled with a set of rules. The left hand side is a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label (Medhat *et al*., 2014). There is a research on predicting public emotion based on the posts during crisis on social media (Brynielsson *et al*., 2014). The performance of result that used simple rule-based system was less good than SVM. Another research is on predicting the sentiment on top Indian e-Commerce vendors from social networks (Makaram *et al*., 2015). The data is collected from Twitter for 5 vendors. The result showed that flipkart is more famous with highest positive and lowest negative feedbacks. It helps the vendors to compete with good strategies (Makaram *et al*., 2015).

Probabilistic classifier: Probabilistic classifiers use mixture models for classification. The mixture model assumes each class is a component of the mixture where each of them is a generative model that provides the probability of sampling a particular term for component. There are three main classifiers include Naive Bayes (NB), Bayesian Network (BN) and Maximum Entropy (ME) (Medhat *et al*., 2014). NB is the simplest and commonly used classifier. This classification model computes the posterior probability of a class based on the distribution of the words in the document (Medhat *et al*., 2014). Furthermore, this classifier is used in a research related to online Malay posts (Alsaffar and Omar, 2014). Besides, this classifier is also used in formal writing and informal writing styles (Kaur and Saini, 2014). In Fig. 2, NB scored the highest accuracy on informal writing style, which used different languages.

BN is a directed acyclic graph, which the nodes represent random variables and the edges represent conditional dependencies. It is considered as a complete model for the variables and their relationships and specified for a complete Joint Probability Distribution (JPD) over all the variables (Medhat *et al*., 2014). Maximum Entropy (ME) is a classifier that encodes the labelled feature sets to vectors. It is normally used to find the possible label for feature set by conducting weight calculation for each feature and combined the result (Medhat *et al*., 2014). There is a research applied 2 classifiers to classify sentiment of data from Twitter in predicting of U.S elections. It has used Latent Dirichlet Allocation (LDA) algorithm to extract the famous topics in discussions (Jahanbakhsh and Moon, 2014). Two observations are conducted on 2 nominations, obama and Romney. In second observation, Obama lead the popularity with a large gap than Romney. This research succeeded on predicting the winning candidate for that election.

MATERIALS AND METHODS

Unsupervised learning: Unsupervised learning approach uses for classifying the unlabeled data. Many researchers used this approach for specified levels (Medhat *et al*., 2014). It is widely used since the data is easier to collect for training rather than labelled data (Medhat *et al*., 2014). This approach is used to find aspects in reviewing Chinese social media and the sentiments are expressed in different aspects. It helps in exploring multi-aspect fine grained topics and associated sentiment (Medhat *et al*., 2014). Deep Learning (DL) uses to set levels of hierarchy of features. This learning is designed to use unlabeled data to extract high-level features from reviews for domain adaptation. It has applied on Amazon reviews to classify sentiment polarity. The result shows this approach is not over the baseline of SVM.

Lexicon-based approach: Lexicon-based approach involves opinion words which is divided into 2 main categories. Positive represents desired expression and negative are vice versa (Medhat *et al*., 2014). This approach requires sentiment lexicon to generate it either by manual or semi-automatically. The main approaches in collecting the opinion word list either is manually or semi automatically are dictionary-based and corpus-based (Medhat *et al*., 2014). There are some manual approaches like general inquirer, opinion lexicon and the polarity is assigned by human Muhammad. These approaches is very time consuming and must used together with the automated approaches to avoid the mistakes in the result (Medhat *et al*., 2014).

Dictionary-based approach: This approach uses for small set of opinion words where it is collected manually with known orientations. Then the set is extended by searching its synonyms and antonyms in the known corpora or thesaurus. This process will be iterated until no new word is found. After that, the manual inspection is carried out to remove or correct errors (Medhat *et al*., 2014). There is application generates sentiment analysis on political reviews such LIWC Software that uses dictionary-based. This software is used by a lead curator of Research Associate from Western Michigan University to conduct a research on measuring variables in conversation between the respondents. Basically, it uses for psychology and linguistics tasks on measuring the sentiment levels in US Senatorial speech.

Corpus-based approach: Corpus-based approach uses to find the opinion words with context specific orientations problems. The solution is relying on patterns that occur with a list of seed of opinion words. It is used on a large corpus (Medhat *et al*., 2014). There are 2 main approaches include statistical and semantic.

Statistical approach is used to find a pattern of co-occurrence. This approach derives polarities of adjectives occurrence in a corpus. It is done by using the set of indexed documents on the web as corpus for the construction of dictionary. In addition, it is able to handle the unavailability of some words because the size of corpus is not large enough. The identification polarity of a word is based on the occurrence frequency of the word within corpus (Medhat *et al*., 2014). Semantic approach directly gives sentiment values and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantic close words as example, WordNet. It computes the sentiment polarities (Medhat *et al*., 2014). There is a research used lexicon-based approach to do sentiment analysis on the word in Malay and English post from Facebook page. This research tags with emotion polarity classes as happy, unhappy or emotionless (Zamani *et al*., 2013; Khan *et al*., 2015). Happy represents positive, unhappy represents negative and emotionless represents neutral opinion word.

RESULTS AND DISCUSSION

Hybridization approach: Some researches use hybrid approaches which involves various types of classification approaches. There is a research that compares and combines the methods of lexicon based such as SentiWordNet, LIWC, happiness index and etc. It analysed the posts based on major topics such as Olympics and H1N1 from Twitter and other social media (Goncalves *et al*., 2013). Two other approaches, dictionary-based and SVM are used to do sentiment classification on polarity for labelling of training data according to domain application on entity level. The data used is from Twitter. The result of research discovers the improvement on precision and call over baseline. Another research is on concept level. The implementation uses pSenti of lexicon-based and SVM to sentiment the software and movie reviews from Twitter. On sentiment the polarity classification, it scores high accuracy than the pure learning-based and nearest to the pure-learning based (Mudinas *et al*., 2012).

There is another research applied 3 sentiment analysis approaches, which are rule based, lexicon-based and machine learning approaches in pipeline system architecture. Furthermore, SentiStrength of lexicon-based for summation of each word polarity and SVM is used for feature extraction process. The research classified the polarity of expression and message level of Twitter posts (Filho and Pardo, 2013). There is a system that uses hybrid of three modules include dictionary based sentiment analysis, pattern based sentiment analysis and semantic event based sentiment analysis. This research focused on stocks domain to gain accuracy on aspects of technical and linguistic problem. The result on sentiment is represented in graph to show the impact of positive and negative according to events and stock price based on time frame.

Negation is part of opinions. There is research used rule-based to define the rule combining antecedent and consequent if-else relation and SVM for data inspection and pattern identification for classification and regression (Kawathekar and Kshirsagar, 2012). The result found that the rule based approach is more preferable than SVM in sentiment the online movie reviews (Kawathekar and Kshirsagar, 2012). There is another research involves political sentiment on prediction for the election results (Bermingham and Smeaton, 2011). It combines SVM, multinomial NB and volume-based measures. The research evaluated the conventional election polls with the final election result (Bermingham and Smeaton, 2011).

Finally, there is a research that has been conducted on corpus articles of New York times which are convenient for languages that lack of general dictionary resources. It combined 2 process approaches between semi-supervised graph-based algorithm and supervised models which involved three tasks. Based on the evaluation result it is equivalent to the popular lexicon aspects for mono sentimous words only, means a word does not ambiguous (Glavas *et al*., 2012).

After reviewing the literatures this study found there are only few researches that applied sentiment analysis techniques to classify social media posts. Table 1 shows the comparison of the sentiment analysis approaches. The approaches are evaluated based on social media types, classifiers used for experiments and applicability on political issues (Table 1). Based on the comparison table, mostly researches used data from Twitter. It is due to the accessibility of tweets that eases the third party to collect data. In term of type of classifier most of the researches focus on linear classifier and hybrid approach. Additionally, there are researches on hybrid classifiers which used several classifiers of machine learning approaches such as SVM, rule-based, lexicon-based and others.

Seem to be the future research will focus on political domain. Therefore, applicability on political area is one of the important criteria in this comparison. Table 1 shows that only 2 researches used hybrid approaches in sentiment politic related data, which involve hybrid of lexicon dictionary based and SVM and hybrid of volume-based measure and supervised learning respectively. From the literature, there are only 2 researches that applied sentiment analysis approaches on

Table 1: Comparison of approach and classifiers

		Comparison criteria	
Classifier/Approach	Social media types	Algorithm/classifier used	Applied on politic
Linear	Twitter	Multinomial naive bayes, SVM	No
	Twitter	SVM, NN	No
	Website	Dictionary based Sentiment Analysis (SA), pattern based SA, semantic event SA	No
	Website	Semi-supervised and supervised learning	No
	YouTube, Twitter	Combined of lexicon approaches	No
Hybrid	Twitter	Lexicon of dictionary based and SVM	Yes
	Twitter	Lexicon of psenti and linear classifier	No
	Twitter	Lexicon of senti strength and SVM	No
	Twitter	Rule-based and SVM	No
	Twitter	Volume-based measure and supervised learning	Yes

Table 2: Comparison of previous research using Malay posts

				Comparison criteria	
Author	Classifier/Approach	Social media types	Algorithm/ classifier used	Data review	Polarity classes
Alsaffar and Omar (2014)	Linear, Probabilistic	Social media and blog	SVM,NB,KNN	Online Malay written	Positive, negative
Amani	Lexicon-based	Facebook	Lexicon-based	Facebook comment at certain page	Happy, unhappy, emotionless

Malay language posts. Table 2 shows the comparison of 2 previous researches using Malay posts as data. The comparison is based on type of classifier, type of social media data, algorithm, data review and polarity classes. Based on Table 2 these researches used Malay posts from social media include blog and Facebook. Several approaches are used such as Lexicon-based, SVM, NB and KNN. Based on their results, it found that SVM is better than others sentiment analysis approaches in classify the polarity. In term of polarity classes these researches used the sentiment of positive and negative and the emotion of public as happy, unhappy and emotionless. Based on these two comparison tables, the most suitable approach for the future research is the hybrid approach of Lexicon of Dictionary-based and SVM. It is because from the review, it shows that these approaches have been used to classify the sentiment and at the same time used Twitter as data. Besides, these approaches have been proven which applicable in politic area. Nevertheless both approaches have been applied separately on Malay language posts. In the future research, the hybrid of lexicon of dictionary-based and SVM will be used as the classifiers for the sentiment analysis on Malay language Twitter posts that related to political issue. The future research will be on classifying the Malay political tweets and sentiment them with correct polarity. However, the dataset has to be pre-processed according to grammar rules and structures of Malay language.

CONCLUSION

Social media data contains big data which is collected and analyzed everyday to improve the work strategies in business, investment, political policy and national security. One of the crucial data in sentiment analysis is sentiment value that indicates the polarity of the social media postings. Moreover, it also represents the opinions of public about hot topics in the media. There are many researches utilized sentiment analysis approaches to help to sentiment the polarity of sentences or postings. These researches are reviewed and compared in this study based on the type of classifiers, social media types, algorithm used and the applicability in politic domain and also language used in postings. This study found the suitable approach for future work which is the hybrid of lexicon dictionary-based and support vector machine for classifying the sentiment polarity of Malay tweets for political inclination.

ACKNOWLEDGEMENTS

The researcher express appreciation to Universiti Teknologi MARA (UiTM), Malaysia through Lestari Research Grant (600-IRMI/DANA 5/3/LESTARI (0125/2016)) for sponsoring this study.

REFERENCES

Alsaffar, A. and N. Omar, 2014. Study on feature selection and machine learning algorithms for Malay sentiment classification. Proceedings of the International Conference on Information Technology and Multimedia (ICIMU), November 18-20, 2014, IEEE, New York, USA., ISBN:978-1-4799-5423-0, pp: 270-275.

Bermingham, A. and A.F. Smeaton, 2011. On using twitter to monitor political sentiment and predict election results. Proceedings of the Workshop on Sentiment Analysis where AI Meets Psychology, November 13-13, 2011, Dublin City University, Dublin, Republic of Ireland, pp: 2-10.

Brynielsson, J., F. Johansson, C. Jonsson and A. Westling, 2014. Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. Secur. Inf., 3: 1-11.

Dickinson, B., 2015. Detecting Illegal Drug usage in Social Media using Support Vector Machines and Deep Neural Networks. University of Rochester, Rochester, New York,.

Filho, B.P.P. and T.A. Pardo, 2013. Nilc usp: A hybrid system for sentiment analysis in Twitter messages. Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics, June 14-15, 2013, University of São Paulo, São Paulo, Brazil, pp: 568-572.

Glavas, G., J. Snajder and B.D. Basic, 2012. Experiments on hybrid corpus-based sentiment lexicon acquisition. Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, April 23-23, 2012, ACM., Stroudsburg, USA., pp: 1-9.

Goncalves, P., M. Araujo, F. Benevenuto and M. Cha, 2013. Comparing and combining sentiment analysis methods. Proceedings of the 1st ACM Conference on Online Social Networks, October 07-08, 2013, ACM, Boston, Massachusetts, ISBN:978-1-4503-2084-9, pp: 27-38.

Jahanbakhsh, K. and Y. Moon, 2014. The predictive power of social media: On the predictability of U.S. president elections using Twitter. Comput. Sci. Soc. Inf. Networks, 2014: 1-10.

Kaur, J. and J.R. Saini, 2014. Emotion detection and sentiment analysis in text corpus: A differential study with informal and formal writing styles. Int. J. Comput. Applic., 101: 1-9.

Kawathekar, S.A. and M.M. Kshirsagar, 2012. Sentiments analysis using hybrid approach involving rule-based and support vector machines methods. IOSRJEN., 2: 55-58.

Khan, A.Z., M. Atique and V.M. Thakare, 2015. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Int. J. Electron. Commun. Soft Comput. Sci. Eng., 2015: 89-91.

Makaram, D.R.A., A. Sridhara and F.J. Angeline, 2015. Rule based classifier to auspicate the sentiment towards the top Indian E-commerce vendors through social networks. Int. J. Comput. Technol. Appl., 6: 431-439.

Medhat, W., Hassan, A. and H. Korashy, 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J., 5: 1093-1113.

Mudinas, A., D. Zhang and M. Levene, 2012. Combining lexicon and learning based approaches for concept-level sentiment analysis. Proceedings of the 1st International Workshop on Issues of Sentiment Discovery and Opinion Mining, August 12, 2012, ACM, New York, USA., ISBN:978-1-4503-1543-2, pp: 1-5.

Nirmala, K., S.S. Kumar and D.J.A. Vellingiri, 2013. Survey on text categorization in online social networks. Int. J. Emerging Technol. Adv. Eng., 3: 446-450.

Pandhe, S. and S. Pawar, 2015. Algorithm to monitor suspicious activity on social networking sites using data mining techniques. Int. J. Comput. Appl., 116: 35-40.

Patel, M.P. and M.K. Mistry, 2015. A review: Text classification on social media data. IOSR. J. Comput. Eng., 1: 80-84.

Zamani, N.A.M., S.Z. Abidin, N.A.S.I.R.O.H. Omar and M.Z.Z. Abiden, 2013. Sentiment analysis: Determining people's emotions in Facebook. Appl. Comput. Sci., 2014: 111-116.