IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 9, No. 3, September 2020, pp. 473~479
ISSN: 2252-8938, DOI: 10.11591/ijai.v9.i3.pp473-479

Machine learning-based technique for big data sentiments extraction

Noraini Seman, Nurul Atiqah Razmi
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received Feb 15, 2020
Revised Apr 24, 2020
Accepted May 20, 2020

Keywords:

Classifier
F-measure
Machine learning
Recall
Sentiment analysis

ABSTRACT

A huge amount of data is generated every minute for social networking and content sharing via Social media sites that can be in a form of structured, unstructured or semi-structured data. One of the largest used social media sites is Twitter, where each and every day millions of data generated in the form of unstructured tweets. Tweets or opinions of the people can be used to extract sentiments of the people. Sentiment analysis is beneficial for organizations to improve their products and make required changes on demand to increase their profit. In this paper, three machine learning algorithms Support Vector Machine (SVM), Decision Trees (DT), and Naive Bayes (NB) for classifying sentiments of twitters data. The purpose of this research is to compare the outcomes of these algorithms to identify best machine learning method which gives most accurate and efficient results for classifying twitter data. Our experimental result shows that same preprocessing methods on a different dataset affect similarly the classifiers performance. After analyzing the results it is observed that SVM provides 64.96%, 71.26% and 91.25% precision which is better than other two algorithms. Also, overall Recall and F-measure rate of SVM is greater than NB and DT for three datasets. However, it is important to further study current available preprocessing techniques that help us to improve results of various classifiers.

Corresponding Author:

Noraini Seman,
Department of Computer Science,
Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA, Shah Alam, Malaysia.
Email: aini@tmsk.uitm.edu.my

1. INTRODUCTION

Twitter, the largest used social media site, has now become a very popular trend over the world for people who want to share an opinion about their social, political and economic interest. User opinion can be related to various aspects like gadgets, politics, products, services etc. that can directly convey the viewpoint of the user and helps in making predictions of a consumer market. Such kind of opinions or sentiments of huge people around the world is capable of performing analysis and future predictions. Usually, tweets contain incomplete, poorly structured, noisy, irregular expressions, ill-formed words and non-dictionary terms [1]. Also, messages or tweets are short and have 140 lengths of limitations. So it requires preprocessing done on our collected datasets to reduce noise in tweets by removing stop-words, removing URLs, replacing negations etc [2]. Sentiment dictionary contains all forms of a word with each word's polarity strength that can save more time.

Sentiment analysis (SA) is a process of detecting the contextual polarity of text in terms of positive, negative or neutral [3-4]. Organizations across the world widely adopted the ability to extract insights from these sentiments of various social media sites. It helps organizations to make predictions of a certain product, reviews, and other decision-making processes that will ultimately increase the profit. So ultimately SA is beneficial for organizations and individuals to improve their profit as per user or market demand. SA also known as opinion mining, is a most popular trend in today's world which is the process of identifying and categorizing opinions on the web, determines the writer attitude towards a particular topic or product [5]. It tells about what author wants to communicate and defines his state of mind in terms of emotions, feelings, and subjectivities about an event or topic. It involved with Natural Language Processing (NLP) process which is the interaction between the computers and the human/natural language [6-8]. NLP technique facilitates easy pre-processing of text i.e. NLP cleans and normalizes text for sentiment analysis [8]. Analysis of sentiments can be based on single phrase or sentence, where the sentiment of the whole sentence is calculated. It contains following steps [9-10]:

- Tweets posted on twitter are freely available through a set of APIs of twitter. At first, we collected a corpus of positive, negative, neutral and irrelevant tweets from twitter API.
- Then pre-processing done by removing stop words, negations, URL, full stop, commas etc. to reduce noise from tweets and to prepare our data for sentiment classification.
- Then, we apply machine learning algorithms to our dataset and compare their results.
- Results help us to identify which machine learning algorithm is best suited for classification of SA.

Applications of SA are broad and powerful that provide us easier and quicker social media monitoring like in: Consumer market for product reviews; Marketing to know consumer trends and attitude; Social media to find general user opinion about current topics; Movie to know whether released movie is liked or not, etc [11]. As users on social media sites are rapidly growing and producing a large amount of data every day, so there is a need to classify and analyze these messages to find out its polarity about some topic or event [12-13]. Emotions and opinions can be expressed in many ways. Classifying sentiments that have few relative classes such as "positive", "negative", or "neutral", is the most complicated task. SA is a popular topic and lots of research has been going on from a long time. Many researchers used supervised learning algorithms also with various automatic classifiers for classification of the polarity of sentiments [14]. The problem is in assigning the strongest polarity of sentiments and in finding the best algorithm which provides most accurate results.

In this paper we use three machine learning algorithms Support Vector Machine (SVM), Decision Tree (DT) [15] and Naïve Bayes Classifier (NB) sentiment classifier for classifying our data also helps in evaluating the performance of our training dataset. We focused on comparing outcomes of these algorithms to identify best machine learning method which gives most accurate and efficient results for classifying twitter data.

2. RESEARCH METHOD

This paper presents a model presented in Figure 1, which consists of three layers for analyzing sentiments. First Data Collection layer, used to collect tweets from twitter APIs; Second Data preprocessing layer with a selection of attributes which is used to reduce noise level from tweets, and last SA or Data Mining layer used to apply machine learning algorithm [2].

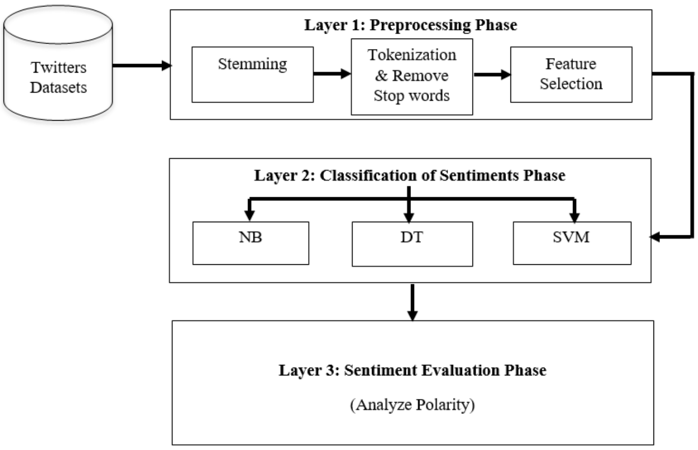
Twitters Datasets
Layer 1: Preprocessing Phase
Stemming
Tokenization & Remove Stop words
Feature Selection
Layer 2: Classification of Sentiments Phase
NB
DT
SVM
Layer 3: Sentiment Evaluation Phase
(Analyze Polarity)

Figure 1. General model of sentiment analysis

2.1. Data collection

At first, we obtain training data of twitter sentiments from 2- different twitter API. First, dataset taken from “Twitter Sentiment System for SemEval 2016”, (denoted by “SE-T”) contains approx 13541 tweets with 2-attributes namely: class and content [16]. Second dataset is taken from “Sanders Analytics twitter sentiment corpus” (denoted by TS), which contains 479 instances with class and text two attributes [17] as presented in Table 1. However, we also collect our own twitter data in Malay language, which is spoken in Malaysia, Singapore, Indonesia, and a few other countries and denoted as “OC”. This language is actually the fourth-most popular language on Twitter, accounting for 8 percent of all Tweets are about airlines [18].

Table 1. Twitter data collection

Dataset	Total Tweets	Positive	Neutral	Negative
E-Twitter (SE-T)[16]	13541	5232	6242	2067
Twitter Sanders (TS)[17]	479	163	-	316
Own Collection (OC)	6459	3323	1190	1946

2.2. Featurization

Features in machine learning is basically numerical attributes from which anyone can perform some mathematical operation such as matrix factorization, dot product etc. But there are various scenario when dataset does not contain numerical attribute for example- sentimental analysis of Twitter/Facebook user, Amazon customer review, IMDB/Netflix movie recommendation. In all the above cases dataset contain numerical value, string value, character value, categorical value, connection (one user connected to another user). Conversion of these types of feature into numerical feature is called futurization.

2.2.1. Text processing setup

For the purpose of getting accurate results by classifiers we have to make sure that these datasets processed efficiently by removing unrelated contents and thus related contents are accurately extracted. As most researchers consider that URL doesn’t have any information regarding sentiments, so by removing short URLs from tweet contents can be refined. People often use emotional words that contain repeated letters to express their sentiments which are very common trends like “coooool”. Also, numbers are not used for analyzing sentiments so tweet contents can be refined by removing them [1]. The polarity of the word will be changed when they are preceded by a negation or negation can change/reverse the meaning of words. By checking negations, Removing of URLs, emotions, numbers and Repeated Word; noise in tweets can be reduced. This filter provides us options to do configuration with our dataset which includes following steps [19-20]:

- Stemming: It is used to remove suffix from the word according to some grammatical rules. Here we apply most popular Snowball Stemming library.
- Stop Word Extractor: Some words that don’t have polarity so they don’t need to be further analyzed like: able, are, both, which, has, become, after etc. So after elimination of these words, our result will not be affected. We used Rainbow list for our experiment.
- Tokenization: It is used to split a document into a word or terms and make a word vector. Here we used NGramTokenizer.
- Feature Selection: This process decreases the number of attributes into a better subset which can increase accuracy also it brings a reduction in training time. It is done by using Filters and Wrappers.

2.3. Sentiment classifier

To classify sentiments machine learning (ML) algorithms are used i.e. a branch of Artificial Intelligence (AI) concerned with the study of classification and pattern analysis, allows the computer to learn behaviors of empirical data taken from sensors or database [21]. ML algorithm allows us to automatically recognize complex patterns and make intelligent decisions based on data. In this paper, we used various machine learning algorithms such as Naive Bayes (NB), Support Vector Machine (SVM) [22], and Decision Tree (DT) [15].

2.3.1. Naïve bayes classifier

It refers to counting the frequency of words that are related to the sentiments in the message. As Bayes theorem based on probabilistic classifier so it allows us to capture uncertainty about the model to determine the probability of the outcome. Explicit probabilities can be calculated by it for the tested dataset and it helps to reduce noise robustly. It is numerical based approach with easy, fast and high accuracy features.

2.3.2. Support vector machine (SVM)

It yields more accurate results when it is used for classifying text. The basic idea behind it is to find the hyperplane (or vector w), which is responsible for separating one class document vector from the vector in other class [7]. It is successfully employed in text classification and various other sequence processing applications as it is a type of linear classifier.

2.3.3. Decision tree (DT)

It is a flowchart used to output labels for certain features, act as input values. It categories a document as by, starting from the tree root (labeled as features), followed downward by branches (labeled as features weight) and last reached a leaf node (labeled by categories).

2.4. Experimental setup

We use Waikato Environment for Knowledge Analysis (WEKA) to implement data mining algorithms for preprocessing, classification, clustering, and analysis of results [23-24]. This environment includes java libraries that implement algorithms and provide the best environment to researchers for classifying datasets. We apply "StringToWordVector" filter and done lots of preprocessing with our datasets [25-26]. Using n-gram tokenizer option and attribute selection method different number of attributes are created. With attributes selection method 50 attributes are taken for testing out of 1613 words from first dataset SE-T [16] and 105 attributes out of 2065 words are taken from second dataset TS [17]. This method increases accuracy rate of our training dataset also it brings a reduction in execution time. Following Table 2 shows reduction in size of file after preprocessing:

Table 2. Data collection criteria

Dataset	E-Twitter (SE-T) [15]	Twitter Sanders (TS) [16]	Own Collection (OC)
Size of file before preprocessing	1.7 MB	93.7 KB	1.5 MB
Size of file after preprocessing (feature selection)	181 KB	9.9 KB	150 KB

To evaluate performance we apply 10-fold cross validation technique which splits the original set into training sample to train the model and a test set to evaluate results. For computing sentiments quickly of tweets without compromising accuracy, an approach known as "Information Retrieval Metrics" can be used to evaluate experimental results in terms of precision, recall, f-measure, and accuracy with the use of following formulas [9, 27]:

$$\text{Precision} = TP/\,(TP+ FP) \tag{1}$$

$$\text{Recal} = TP/\,(TP+FP) \tag{2}$$

$$\text{F-measure} = 2^{*}\ \text{Precision}^{*}\ \text{Recall}/\,(\text{Precision}+ \text{Recall}) \tag{3}$$

$$\text{Accuracy} = TP+TN/\,(TP+ TN+ FP+ FN) \tag{4}$$

Here (TP= True Positive; TN= True Negative; FP=False Positive; FN= False Negative)

3. RESULTS AND ANALYSIS

We observed that our classification results improved in terms of time and accuracy using processed and small features data than simple datasets. For example in first SE-T dataset, time taken to build a model for NB algorithm takes 10.56 seconds, accuracy 53.73% and after processing time taken to test model on training data is reduced at 0.35 seconds only, accuracy improved by 57.46%. Table 3 demonstrates the accuracy of classifiers on three datasets after applying various preprocess methods.

Table 3. Accuracy criteria for datasets

Evaluation Criteria	Dataset	SVM	DT	NB
Correctly classified instances	SE-T	8335	8118	7776
	TS	419	412	355
	OC	5334	5035	5324
Incorrectly classified instances	SE-T	5206	5423	5765
	TS	60	67	124
	OC	1125	1424	1135
Accuracy (%)	SE-T	61.55	59.95	57.42
	TS	87.47	86.01	74.11
	OC	82.58	77.95	82.43
Error	SE-T	0.38	0.40	0.42
	TS	0.13	0.14	0.26
	OC	0.17	0.22	0.18

Following performance measures are reported in Table 4 by our experimental result using three dataset, after conducting 10-fold cross validation technique.

Table 4. Performance measures of classifiers

Classifier	Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	Polarity
SVM	SE-T	0.352	0.036	0.859	0.352	0.500	positive
	TS	0.906	0.562	0.580	0.906	0.707	neutral
	OC	0.644	0.006	0.981	0.644	0.778	positive
DT	SE-T	0.34	0.048	0.821	0.349	0.489	positive
	TS	0.284	0.054	0.486	0.284	0.359	negative
	OC	0.994	0.399	0.828	0.994	0.904	negative
NB	SE-T	0.780	0.479	0.582	0.780	0.667	neutral
	TS	0.252	0.006	0.953	0.252	0.398	positive
	OC	0.437	0.186	0.597	0.437	0.505	positive

The number of correctly classified instances and accuracy rate is greater for three datasets with SVM algorithm. In our experiment obtained accuracy using SVM algorithm is 61.55%, 87.47% and 82.58% respectively (with 50 feature SE-T, 105 feature TS datasets and 100 feature of OC datasets) which is greater than other two algorithms. Our experimental result shows that same preprocessing methods on a different dataset affect similarly the classifiers performance. After analyzing results of Table 4 it is observed that SVM provides 64.96%, 71.26% and 91.25% overall precision which is better than other two algorithms. Also, overall Recall and F-measure rate of SVM is greater than NB and DT for three datasets. Furthermore, time taken to build a model is greatly reduced by applying feature selection method. Time taken to build model in first SE-T datasets is 0.45, 29.43, 4.47 seconds respectively with NB, SVM, and DT algorithm; in second TS dataset, it is 0.01, 0.06, 0.01 seconds with NB, SVM and DT algorithms respectively.

4. CONCLUSION

In this paper, we discuss sentiment analysis which can tell us the thought of writers about the particular entity. These days, it becomes a routine task to find people sentiments about a real world entity from social media sites like Twitter, face book or blogs etc. To efficiently analyze this large amount of datasets it is essential to accurately classify it. In this paper, we have presented a methodology of text mining using Weka tool for classifying sentiments of twitter. We use three machine learning algorithms SVM, DT, and NB for classifying sentiments of twitters data. We conduct an experiment on three twitter’s datasets to verify the effectiveness of pre-processing. Our experimental results indicate that by removing unwanted words and selecting features in the preliminary phase of preprocessing, time to build model is reduced and also it provides more accurate results in applied algorithms. The result may be affected by the choice of features for training and choice of algorithm for sentiment classification. The performance of SVM, DT, and NB algorithms improve on datasets after removing unwanted words. Therefore, removing unwanted words is useful to improve the performance of sentiment classification. We discuss the comparative analysis of three algorithms and calculate overall performance measures in terms of precision, recall, and f-measure. Our experimental results indicate that SVM provides more accurate results than other algorithms. However, it is important to further study current available preprocessing techniques that help us to improve results of various classifiers. A method should be found to automatic incorporate feature selection at time of model building according to any language.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknologi MARA (UiTM) for the research funding and support via grant number 600-IRMI/PERDANA 5/3 BESTARI (107/2018).

REFERENCES

[1] J. Zhao, G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," *IEEE Access*, DOI 10.1109/ACCESS. 2017, 2672677.

[2] A. Krouska et al., "The effect of preprocessing techniques on Twitter SA," in *7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Research Gate Conference: 2016.

[3] B. D. Savita, et al, "Sentiment Analysis on Twitter Data Using Support Vector Machine," in *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 4(3), 2016.

[4] Hemalatha, *et al*, "Sentiment Analysis Tool using Machine Learning Algorithms," in *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 2(2), 2013.

[5] G. Vinodhini, *et al*, "Sentiment Analysis and Opinion Mining: A Survey," in *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 2(6), 2012.

[6] A. Pak et al, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Universite de Paris-Sud, Laboratoire LIMSI-CNRS, FRANCE, pp. 1320-1326.

[7] M. Rani et al, "A Review of Data Analysis of Twitter," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6(5), 2016.

[8] L. Barbosa, *et al*, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," Coling 2010: Poster, pp. 36-44, Beijing, 2010

[9] B. Pang, *et al*, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proceedings of EMNLP 2002*, pp. 79-86, 2002.

[10] H. Anber, *et al*, "A Literature Review on Twitter Data Analysis", *International Journal of Computer and Electrical Engineering*, vol.5 (3), pp: 53-60, 2016.

[11] S. A. Mulay, *et al*, "Sentiment Analysis and Opinion Mining With Social Networking for Predicting Box Office Collection of Movie," in *International Journal of Emerging Research in Management &Technology*, vol. 5(1), pp. 227-235.

[12] Y. Yengi, *et al*., "Distributed Recommender Systems with Sentiment Analysis", *European Journal of Science and Technology*, vol. 4(7), pp. 51-57.

[13] S. Wakade, *et al*, "Text Mining for Sentiment Analysis of Twitter Data", The University of Akron, Department of Computer Science.

[14] R. Nivedha and N. Sairam, "A Machine Learning based Classification or Social Media Messages", *Indian Journal of Science and Technology*, vol 8(16), pp. 102-110.

[15] H. Shamsudin, *et al*, "Hybridisation of RF(Xgb) To Improve The Tree-based Algorithms in Learning Style Prediction," in *IAES International Journal of Artificial Intelligence* (*IJ-AI*), vol. 8(4), pp. 422-428, 2019.

[16] W. Sidorenko, "SemEval-2016 Task 4: Sentiment Analysis on Twitter, Training + Dev dataset," *https://github.com/WladimirSidorenko/SemEval-2016*.

[17] S. Sanders, "Sanders Analytics twitter sentiment corpus," *https://github.com/guyz/twitter-sentimentdataset*.

[18] M. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," vol. 6(6), pp. 1067-1073, 2016.

[19] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," *Int. J. Adv. Sci. Eng. Inf. Technol*., vol. 6, no. 6, pp. 1148-1153, 2016.

[20] M. S. Neethu, and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques." *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013*.

[21] M. A. Al-Hagery, "Extracting Hidden Patterns From Dates' Product Data Using A Machine Learning Technique, " in *IAES International Journal of Artificial Intelligence* (*IJ-AI*), vol. 8(3), pp. 205-214, 2019.

[22] S. Ibrahim, *et al*, "Rice Grain Classification Using Multi-Class Support Vector Machine (SVM)," in *IAES International Journal of Artificial Intelligence* (*IJ-AI*), vol. 8(3), pp. 215-220, 2019.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "*The WEKA Data Mining Software: An Update; SIGKDD Explorations*," vol.11 (1), 2009.

[24] R. Arora and S. Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," *International Journals Computer Application*, vol. 54(13), pp. 21-25, 2012.

[25] N. Mallios, E. Papageorgiou, M. Samarinas, and K. Skriapas, "Comparison of machine learning techniques using the WEKA environment for prostate cancer therapy plan," in *Proceedings of the 2011 20th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2011*, pp. 151-155, 2011.

[26] T. Garg and S. S. Khurana, "Comparison of classification techniques for intrusion detection dataset using WEKA," *Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2014*, 2014.

[27] B. M. Patil, D. Toshniwal, and R. C. Joshi, "Predicting burn patient survivability using decision tree in WEKA environment," *2009 IEEE Int. Adv. Comput. Conf. IACC 2009*, March, pp. 1353-1356, 2009.

BIOGRAPHIES OF AUTHORS

Noraini Seman received the Bachelor degree in Computer Science from Universiti Putra Malaysia (UPM), Malaysia in 1999; the MSc degree from Queensland University of Technology (QUT), Australia (2002), and the Ph.D. degree from the Universiti Teknologi MARA (UiTM), Malaysia (2012). She is presently Head of Academic Program Accreditation under Curriculum Affairs Unit and senior lecturer at Department of Computer Science, Faculty of Computer and Mathematical Sciences. Her research interests include AI application to digital signal processing problems, speech summarization, machine translation and machine learning techniques in speech recognition technology and text mining with vast techniques/approaches of Natural Language Processing (NLP).

Nurul Atiqah Binti Razmi obtained her Bachelor of Engineering (Mechatronics) in 2012 from Monash University Malaysia. She has been working in Manufacturing Quality in rare-earth products industry for 5 years before pursuing her master’s degree. She has completed her Master of Data Science study at UiTM Shah Alam in January 2020 and is now working in software industry to develop Artificial Intelligence products for banking sector.