Twitter sentiment analysis with word2vec

NguyễnThịThúyHoài
LêThiệnNhậtQuang

***Abstract*—Word2vec has been populated features for text classification tasks such as sentiment analysis. Recently, there are many available word2vec models such as GoogleNews-vectors-negative300 and word2vec-twitter-model that help researchers doing sentiment analysis easier. In this paper, authors proposed new method to solve text classification tasks. We will extract features of tweets (verb or adjective), then run them with available word2vec models before classify the tweets with machine-learning technologies. This aims to provide an effective method for new researchers about sentiment analysis.**

***Index Terms*—sentiment analysis; word2vec; feature attraction; semantics**

I. INTRODUCTION

Twitter has provided the huge tweets that related to information about what people think and feel about their products and services. Nowadays, many companies are extremely interested in Twitter sentiment analysis in order to collect and summary their user's ideas about their products and services. Thus, Twitter sentiment analysis has become one of interesting domain for information sciences. The goal of Twitter sentiment analysis is not to identity topics, entities, or authors of a text but to automatically identifying sentiment towards products, movies, politicians, etc., as positive, negative, or neutral, rate the expressed sentiment as well as improving customer relation models, detecting happiness and well-being, and improving automatic dialogue systems. Almost tweets are short informal textual messages due to limited in length, usually span one sentence or less, misspell, slang terms and shortened forms of words. Thus, this has brought in many challenges to sentiment analysis [5]:

+ Message Polarity Classification

Given a tweet, predict whether the tweet is of positive, negative, or neutral sentiment

+ Tweet classification according to a two-point scale

Given a tweet known to be about a given topic, classify whether the tweet conveys a positive or a negative sentiment towards the topic

+ Tweet classification according to a five-point scale (VeryPositive + Positive + OK + Negative + VeryNegative)

Given a tweet known to be about a given topic, estimate the sentiment conveyed by the tweet towards the topic on a five-point scale.

+ Tweet quantification according to a two-point scale

Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the Positive and Negative classes.

+ Tweet quantification according to a five-point scale

Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the five classes of a five-point scale.

Recently, there are many different techniques to classify the tweets such as positive, negative or neutral. Some effective machine-learning techniques such as Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes that rely on the features used in the classification task have resolved text classification problem. Several features have been used for this task such as the bag-of-words (BoW), lexical, syntactic, word2vec, doc2vec features [1]. Inside, word2vec feature based on semantic relation between words for several text classification tasks [1].

In this paper, the authors proposed a new method which uses word2vec models (GoogleNews-vectors-negative300 that learned over 3 billion running words) for twitter classification task.

II. RELATED WORK

SA has many traditional algorithms for text classification such as Naïve Bayes, K-means/KNN, and Decision Trees [3].

Mikolow et al. proposed the Feedforward Neutral Net Language Model (NNLM), Recurrent Neutral Net Language Model (RNNLM), Parallel Training of Neutral Networks, and new log-linear models such as CBOW and SG for continuous vector representations of large datasets. Their experiments are resulted that one could train high quality vectors on simple model architectures, such as CBOW or SG.

Several different learning techniques and algorithms for sentiment classification on Twitter data are analyzed and compared by Kharde and Sonawane. They also compared the performance of machine learning algorithms such as Naïve Bayes, Max Entropy, Support Vector Machines and other lexicon-based approaches.Pang and Lee'sexperiment with the support vector machine method achieved the highest accuracy with a classification prediction rate of 86% anda number of other algorithms all in the 80% accuracy range.

Maas et al. introduced a model that combines bothsupervised and unsupervised sentiment components to predictdocument-level sentiment. They were able tocorrectly classify 88% of the test cases that based on widelytesting

***Nguyen ThiThuyHoai,** The University ofDanang, University of Technology and Education, Danang, Vietnam*
***Le ThienNhatQuang,** The University ofDanang, University of Technology and Education, Danang, Vietnam*

corpora and were able to out-perform all of the othermethods previously tested.

The work of Go et al., removed all emoticon and non-word tokens as they train theiralgorithms and they achieved correctly 80% of test cases with Naive Bayes, Maximum Entropy, andSupport Vector Machine algorithms and 83% of ones with the Maximum Entropyclassifier when using both Unigrams and Bigrams

*Kouloumpis et al.*built a dataset that included hashtags (e.g.#bestfeeling, #epicfail, #news) andtested the inclusion of emoticonsagainst the dataset of hashtags alone. Their experimental results showed that includingemoticons showed no significant improvement versus hashtagsalone.*Lilleberg and Yun*used word2vec model for textclassificationagainst tf-iZdf, termfrequency-inverse document frequency, and were able to showthat word2vec in conjunction with tf-idf outperforms tf-idf onits own. Theyachieved extremely effective, 85% accuracy.

Wang et al. compared the efficiency of the TF-IDF algorithm,and the TF algorithm for text classification against the featurebased six-tuple vector model in conjunction with the HighAdverb of Degree Count (HADC) on Chinese sentimentanalysis. They succeeded in classifyingbetween 88% and 92% of reviews accurately.

After that, Mikolov et al. introduced several extensions of the original Skip-gram model named Word2vec. This model not only trained models on several orders of magnitude more data than the previously published models, but also improved the quality of the learned word and phrase representations, especially for the rare entities. Word2vec encompasses two models: the continuous bag-of-words model (CBOW) and skip-gram (SG) model. The CBOW method is used to predict the context by the context, while the SG method uses the word to predict the context.

Villegas *et al*[4] analyzed distributional representations such as BoW, Second Order Attributes (SOS), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Document Occurrence Representation(DOR)and a distributed representation such as Word2vec.Their experiment with LDA and DOR did not perform well, otherwise, word2vec showed good results with the LibLINEAR classifier using default parameters.

EissaM.Alshari et al. [1] proposed feature extraction method based on clustering for word2vec. This method consist of three main components: (1) the discovery of word embedding based on Word2vec, (2) the clustering of term in vocabulary based on opinion words and (3) the construction of features matrix for classification based on cluster centroids. This method reduced the size of the Word2vec feature set for sentiment analysis and constructs cluster of term centered by a set of opinion words from a sentiment lexical dictionary, as well as redistributes the terms in the space based on their polarity.

III. PROPOSED METHOD FOR SENTIMENT ANALYSIS

From knowledge of sentiment analysis and word2vec, author proposes new method in text classification. This method will execute three steps as above image.

Step1: in this word, authors use a data provided for the SemEval-2013 competition: Sentiment Analysis in Twitter (Wilson et al., 2013). The Semeval dataset consists of 825 negative, 2969 neutral and 2216 positive tweets for training and testing. In this step, Semeval dataset is processed in order to get their features (verb or adjective). There are two feature sets after this step (verb feature set and adjective one). Author divides each feature set into training and test set. Particularly, training set consists of 16000 tweets and test set will be the remained one. The target of this work is to detect whether the tweets conveys a positive, negative, or neutral sentiment.

Step2: the set of features that was extracted from step 2 will be processed with GoogleNews-vectors-negative300 model and a feature matrix is constructed after this step.

Step3: At the last step, the feature matrix will be classified with machine learning algorithms. This is one of the most important steps in all machine learning's tasks. The aim of classification is to identify to which set or category a new observation belongs, on the basis of a training set of data containing observations whose class is known. With labeled training set, we only need to choose supervised learning classifiers such as Naïve bayes, logistic regression and support vector machine. Each classifier has its own advantages and disadvantages.

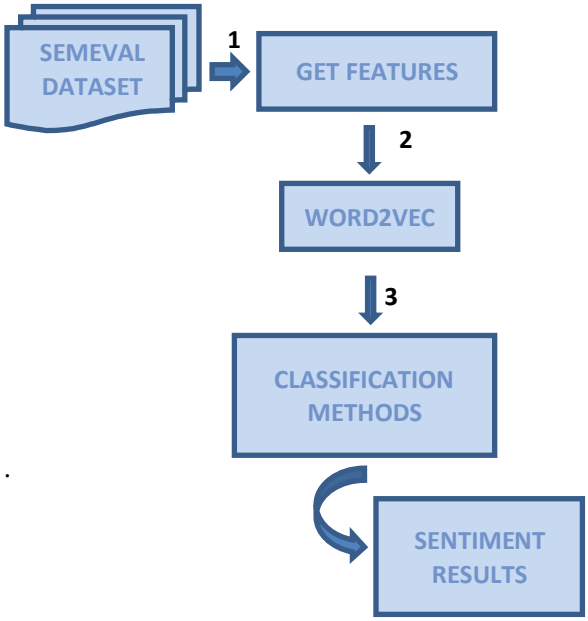

Fig. 1. Framwork for the proposed method

TABLE I. *ADVANTAGES AND DISADVANTAGES OF SUPERVISED LEARNING CLASSIFIERS*

Supervised learning classifiers	**Advantages**	**Disadvantages**
NAIVE BAYES	very low time complexity its assumption usually works quite well in some real-world situations such as spam filtering and document	makes a very strong assumption on the shape of your data distribution data scarcity continuous features

	classification	
SUPPORT VECTOR MACHINE.	it is very effective not only in high dimensional spaces, but also in cases whether the number of dimensions is greater than the number of samples. Second, it is considerably memory efficient due to its own advantage of kernel mapping to high-dimensional feature spaces	if the number of feature is much greater than the number of samples, the method is likely to give poor performance

IV. EXPERIMENTAL RESULTS

Based on the proposed method in the previous section, author executed three experiments for text classification. This set of tweets will train with GoogleNews-vectors-negative300 (3 billion running words).

TABLE II. *ACCURARY OF ADJECTIVE FEATURE SETS FOR SENTIMENT ANALYSIS WITH WORD2VEC*

Tweets (adjective feature)	**GoogleNews-vectors-negative300**
Negative	114/233
Neutral	77/138
Positive	64/324

TABLE III. *ACCURARY OF VERB FEATURE SETS FOR SENTIMENT ANALYSIS WITH WORD2VEC*

Tweets (verb feature)	**GoogleNews-vectors-negative300**
Negative	233/233
Neutral	111/138
Positive	313/324

From table I and table II, it is observed that the proposed method with verb feature has right tweet percent higher than with adjective feature .

V. CONCLUSION

This paper described the concept of sentiment analysis with word2vec model. Besides, author introduced a new method that can help researcher approaching text classification tasks easily. From experimental results in section III, using verb features that are extracted by tweets will bring higher accuracy when they execute with word2vec and classify with Machine learning algorithm.

ACKNOWLEGMENTS

This research is funded by Funds for Science and Technology Development of University of Technology and Education under project numberT2018 – 06 - 92

REFERENCES

[1] Eissa M.Alshri, Azreen Azman, "Improvement of Sentiment Analysis Based on Clustering of Word2vec Features", *International Workshop on Database and Expert Systems Application*, 28 September 2017.

[2] Xiaodan Zhu,Saif M. Mohammad Svetlana Kiritchenko, "Sentiment Analysis of Short Informal Texts," *Journal of Artificial Intellignece Research*, pp. 723-762, Aug. 2014.

[3] Joshua Acosta, NorissaLamaute, Mingxiaoluo, Ezra Finkelstein, AndreeaCotoranu, "Sentiment Analysis of twitter messages using word2vec", *Proceedings of student-faculty research day*, CSIS, Pace University, May 5th 2017

[4] M. P. Villegas, M. Jose, G. Ucelay, J. P. Fern ´ andez, M. A. ´ Alvarez- ´Carmona, M. L. Errecalde, and L. C. Cagnina, "Vector-based word representations for sentiment analysis: a comparative study," pp. 785–793

[5] http://alt.qcri.org/.