International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 2, pp 5-9, February-2021

https://doi.org/10.47001/IRJIET/2021.502002

Semantic Classification Model for Twitter Dataset Using Wordnet

[1]Seham A. Bamatraf, [2]Rasha A. Bin-Thalab

[1,2]Department of Computer Engineering, College of Engineering & Petroleum, Hadhramout University, Mukalla, Yemen

***Abstract* - Twitter is an emerged field in today social media. As twitters increasing, an increasing demand is emerged to mine these twitters and extract useful information. Traditional classification methods have a problem with tweets due to its short sentences. This paper handles the problem of classifying tweets by adapting bag of words feature with semantic tools for natural processing language. The experiments showed a stable performance of classifications in accuracy compared with traditional features of text classification.**

Keywords: Text mining, Classification, Big data, Twitter.

I. INTRODUCTION

Twitter is an online platform that acts as a common blogging tool. It is also considered as a social media site where users share and chat via messages[1]. Accessibility is one of the best things about Twitter. This characteristic makes it so easily for sharing and capturing information. On the other hand, Twitter is one source of big unstructured data like Facebook and other social media. Note that, unlike other media platforms, all twitter messages are public and pullable. This is why Twitter is a gold data miner and has been researched thoroughly to collect useful knowledge. Several trends of twitter research focus on collect[2], analyse its effects [3], and mining[4]. Each domain has several challenges and problems. There are two key problems that prohibit users from gathering useful and realistic social network information. The first is the large volume of shared information that contributes to duplication of details; the second is the rise in confusion and inadvertent disinformation[4].

In this paper we used machine learning algorithms for mining twitters with a special focus on text classification.

The problem with twitter classification in particular due to confusion of short texts and the absence of a large amount of document vector features. Literally, the solution for such issue is accomplished by semantic conceptual incorporation. i.e. extended roles of additional external knowledge or the textual layer of documents with a meaning. This leads to an overload and affects the classification performance. The key study goal here is to enhance traditional features used for text classification. Twitter classification is the task of assigning a set of pre-defined classes to tweets text. A classifier can take this short text as an input, analyse its tokens, and then and automatically assign relevant classes. Like text classification, twitter classification learns to make classification based on previous observation or labels. Text classification is accomplished automatically using machine learning with natural language processing NLP techniques. Two main phases here are followed; training and prediction. The process of text classification is sketched in figure 1.

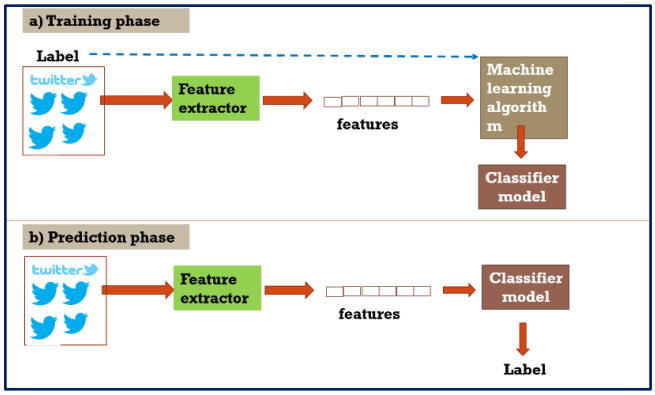

Figure 1: Twitter classification process

In the training phase, the machine learning algorithm is supplied with training data consists of pairs of feature sets and labels to generate a classifier model. Next, in the prediction phase, the classifier model which produced in the training phase, is used to predict labels of new data. The backbone in this process is the features extractor. There are several types of features when developing text classification. Because features are composed of words, we need higher dimensional of features as long as the corpus is broad. Featurization of Text data includes several features such as BOW, TF-IDF[5], word2vector. However, these features often don't take the semanting sense of vocabulary. For the purpose of provide semantic meaning, we need to use additional tools for these features.

In this paper, we focus on provide wordnet with BOW feature. Bag of words (BoW)[6] is one example of features used for text classification. It creates a quantized vector for sentences in documents. In other words, it represents a document with word count and generally ignores the order in which they occur. The vector length is always the same as the

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 2, pp 5-9, February-2021

https://doi.org/10.47001/IRJIET/2021.502002

vocabulary. A long document in which the vocabulary produced is broad will lead to a vector with several zero.

Our contribution in this paper is to adopt BOW feature with WordNet tool to be used more effectively in twitter classification.

The rest of this paper is structured as follows. Section II investigates related work in literature for text similarity and twitter classification. Next, we provide our system approach in section III. Implementation and experiments evaluation are presented in Section IV. Finally, we conclude and address future works in Section V.

II. RELATED WORK

Developed methods in twitter classification can be divided into conventional and semantic methods. Traditional methods focused on vocabulary and neglect the meaning and context of terms or words. Within this area, researchers[7] focused on data training methods for machine learning. Other authors suggested a novel classification method for tweets by utilizing other tweet meta data such as the URLs in the tweet[7], retweeted tweets and tweets from the most influential user of a trend topic. The proposed method which is based on tweet features increases the classification performance in accuracy when compared to traditional approach used like BOW to represent the tweets. Next, additional effort was done based on twitter trending topic classification[8]. The colleague used two different classification approaches: NB Multinomial classifier and C5.0 decision tree learner. Latter, two strategies have been explored[9] in creating paraphrases: WordNet and word embedding which learned from millions of tweets. The authors proved that using word embedding strategy outcomes WordNet dictionary in accuracy. However, word embedding suffers from slow retrieval due to the overhead to search in the embedding information which may reach to 32 Gigabytes like Bablenet [10].

III. BACKGROUND AND PROPSED SYSTEM

In this section we addressed a brief definition and description of WordNet tool and bag of words.

3.1 WordNet

WordNet[11] is an online tool for comprehensive English lexical resource. Nouns, verbs, adjectives and adverbs, each representing a separate term, are grouped into a category of semantic synonyms (called Synsets).Synsets are interconnected through lexical and conceptual relationships. The resultant network of words and definitions can be browsed through a navigator. WordNet framework provides a beneficial method for machine linguistics and natural language processing.

WordNet approaches a thesaurus superficially by ordering words according to their context. This results in a semantically unequivocal shift of words located near to each other in the network. Moreover, WordNet marks the semantic association between words. Super-subordinate association (also called hyperononymy, hyponia or ISA) is the most frequently encoded relationship among synsets. This ties more popular synsets, such as {meets and piece of furniture} to more exclusive ones, such as {bed} and {bunkbed}[11].Thus, a hierarchical is built for all words to connect them semantically. The similarity between two words is measured using path-based similarity. It is a similarity measures to find the shortest number of edges between two synsets utilizing the WordNet hierarchical.

3.2 Bag of Words (BOW)

Bag of Words is a feature used in text classification. Its functionality relies on an algorithm that calculates how frequently a word occurs in a document. As a result, words and documents are listed in a table where rows represent the words and columns represent the documents. Each intersection between row and column determines the number occurrence of a word in a document. Thus each document represented by same length of columns in the corpus. This is called word count vectors which are used later in classification algorithms such as SVM and Random Forest. Let's explain the idea of BOW using an example. Suppose we have three sentences:

```
S1= "Joe waited for the train train"
S2 ="The train was late"
S3 = "Mary and Samantha took the bus"
```

For simplification purpose we eliminate the stop words such as article and auxiliary verbs. BOW includes two steps: determine the vocabulary and count how many occurs of each word for each document. Here, we can suggest each sentence as a document, and then we get the fowling table:

Table 1: Traditional Bag of Words vectors

#	S1	S2	S3
And	0	0	1
Bus	0	0	1
For	1	0	0
Joe	1	0	0
Late	0	1	0
Mary	0	0	1
Samantha	0	0	1
The	0	1	0
Took	0	0	1
Train	2	1	0
waited	1	0	0
was	0	1	0

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 2, pp 5-9, February-2021

https://doi.org/10.47001/IRJIET/2021.502002

IV. METHODOLOGY

The proposed method in this paper is shown in Algorithm 1.

Algorithm 1: Generate Semantic Bag of Words

```
T ← Read twitter sentences
Voc ← Build word vocabulary in the corpus of T
For each sentence ts in T do
Tw ← Extract words from ts
For each word w in Tw
        For each pair in counter and word inVoc
        Bag_vector[counter] =
                Bag_vector[counter]+
              match_score(word,w)
        End For
End For
End For
```

Procedure match_score(w1, w2)

```
If w1== w2
        Match = 1
Else
        Match ← Calculate semantic matching between w1 and
        w2 based on the wordnet path similarity
End If
return Match
```

The algorithm starts by reading all tweets to extract all distinct words. These words formed the vocabulary of the algorithm. Next, we calculate how many times each word in the vocabulary appears for each tweet by using a new function called match_score procedure. The procedure checks two words, if they are same, it returns 1, otherwise it calls WordNet library to compute a similarity path between these two words.

For example, the three sentences in section II. The table computed here will be different, as listed in table 2.

Table 2: Enhanced BOW vocabulary

#	S1	S2	S3
And	0	0	1
Bus	0.67	0.33	1.08
For	1	0	0
Joe	1	0	0
Late	0	1	0
Mary	0.17	0.08	1.08
Samantha	0	0	1
The	0	1	0
Took	0.42	0.58	1.74
Train	2	1	0.42
waited	1.18	0.75	0.44
was	0.52	1.42	0.44

V. EXPERIMENTS

This section defines the experiment that has been performed to evaluate the efficiency of semantic BoW feature in twitter classification. The experiment is conducted using Python which is a general-purpose programming language for data analysis and data science. The primary goal of this experiment is to determine the accuracy of TC. The accuracy is calculated by divided the number of matches by the number of total samples. We begin by presenting the dataset, and then we analyze the results.

5.1 Dataset setup

The twitter dataset used in this experiments is downloaded from [12]. There are 1,578,627 tweets in the Twitter Sentiment Analysis Dataset; each sequence is numbered 1 for positive and 0 for negative sentiment. We used only 600 tweets for simplicity purposes. The dataset is split into two subsets: training and test sets, with a split of 70:30 respectively.

5.2 Results and discussion

To test the performance of our proposed method, we applied six classification methods using traditional BoW and semantic BoW.

We tested semantic BoW using six classification methods include k-neighbors [13], SVM [14], decision tree [15], random forest [16], AdaBoost [17], Multinomial naive Bayes [18].

Table 3: Evaluation the accuracy of classifiers using traditional and semantic BoW

	Semantic BoW	**BoW**
Multinomial NB	69	72
AdaBoost Classifier	69	68
RandomForest Classifier	69	70
DecisionTree Classifier	69	70
SVC accuracy	69	69
KNeighbors Classifier	69	65

It is notable from table 3 to see that the accuracy of semantic BoW is fixed with different classification methods. That is, it accomplished about 72% for the whole methods when the training size is 90% of the dataset size. However, these classification methods typically have different performance according to the quality of the dataset. In addition, each algorithm has limitations and benefits over each other [19]. For example, random forest has best performance over others using traditional features as shown with features BoW+TFIDF.

As shown in table 3, semantic BoW has better performance than traditional features when using with k-neighbors and AdaBoost with 4% and 1% respectively, and same accuracy with SVM. However, Multinominal NB, Decision Tree, and RandomForest has better performance

ISSN (online): 2581-3048
Volume 5, Issue 2, pp 5-9, February-2021
https://doi.org/10.47001/IRJIET/2021.502002

using traditional BoW over semantic BoW by 3%, 1%, and 1% respectively. These classifiers; Multinominal NB, decision tree, and Random Forest, are ideal for discrete features such as word counts. For example, random forest [20] applies a variety of decision tree classifiers and uses the mean to improve predictive accuracy and prevent over-fitting.

VI. CONCLUSION

This paper used the Word Net lexical tool to operate on the Twitter dataset word bag. The method calculates and analyzes the number of words of identical definitions to characterize the tweets. The results demonstrated good performance for improving twitter dataset classification. In future work, we expect to construct a system that will apply and incorporate WordNet with other text features such as TF / IDF to enhance the efficiency of the classification.

REFERENCES

[1] Twitter Inc., 'Twitter turns six', *Twitter turns six*, Mar. 21, 2012. https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html.

[2] R. Szymanski, 'How to Collect Big Data Sets From Twitter'. DZone, Jun. 07, 2019, Accessed: Apr. 12, 2020. [Online]. Available: https://dzone.com/articles/how-to-collect-big-data-from-twitter-for-sentiment.

[3] P. J. Tighe, R. C. Goldsmith, M. Gravenstein, R. Bernard, and R. B. Fillingim, 'The Painful Tweet: Text, Sentiment, and Community Structure Analyses of Tweets Pertaining to Pain', vol. 17, no. 4.

[4] C. Kingston, J. R. C. Nurse, I. Agrafiotis, and A. B. Milich, 'Using semantic clustering to support situation awareness on Twitter: the case of world views', *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 22, Jul. 2018, doi: 10.1186/s13673-018-0145-6.

[5] H. C. Wu, R. W. Pong Luk, K.-F. Wong, and K.-L. Kwok, 'Interpreting TF-IDF term weights as making relevance decisions', *ACM Trans. Inf. Syst.*, vol. 26, no. 13, p. 13:1-13:37, 2008.

[6] Y. Zhang, R. Jin, and Z.-H. Zhou, 'Understanding bag-of-words model: a statistical framework', *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1, pp. 43–52, Dec. 2010, doi: 10.1007/s13042-010-0001-0.

[7] P. Selvaperumal and A. Suruliandi, 'A short message classification algorithm for tweet classification', presented at the Inerantional Conference Recent Trends in Information Technology (ICRTIT), 2014.

[8] A. Zubiaga, D. Spina, V. F. Fernández, and R. Martínez-Unanue, 'Real-Time Classification of Twitter Trends', *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 3, pp. 462–473, 2015.

[9] Q. Li, S. Shah, M. Ghassemi, R. Fang, A. Nourbakhsh, and X. Liu, 'Using Paraphrases to Improve Tweet Classification: Comparing WordNet and Word Embedding Approaches', presented at the IEEE International Conference on Big Data, 2016.

[10] 'BabelNet', *BabelNet*, 2009. https://babelnet.org/about (accessed Apr. 13, 2020).

[11] 'WordNet', *WordNet*, 2005. https://wordnet.princeton.edu/ (accessed Apr. 13, 2020).

[12] 'Twitter Sentiment Analysis Training Corpus (Dataset)', *Twitter Sentiment Analysis Training Corpus (Dataset)*, Sep. 22, 2012. http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/ (accessed Apr. 15, 2020).

[13] N. S. Altman, 'An introduction to kernel and nearest-neighbor nonparametric regression', *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[14] C. Cortes and V. N. Vapnik, 'Support-vector networks', *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[15] L. Rokach and O. Maimon, *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc, 2008.

[16] T. K. Ho, 'Random Decision Forests', in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, 1995, pp. 14–16.

[17] A. and J. Rehman Javed, 'Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles', p. e4088, 2020.

[18] C. D. Manning, P. Raghavan, and H. Schütze, 'Naive Bayes text classification', in *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[19] V. Korde and C. N. Mahender, 'TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY', *Int. J. Artif. Intell. Appl.*, vol. 3, no. 2, pp. 85–99, Mar. 2012.

[20] D. Ignatov and A. Ignatov, 'Decision stream: Cultivating deep decision trees', in *2017 ieee 29th international conference on tools with artificial intelligence (ictai)*, Nov. 2017, pp. 905--912.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 2, pp 5-9, February-2021

https://doi.org/10.47001/IRJIET/2021.502002

Citation of this Article:

Seham A. Bamatraf, Rasha A. Bin-Thalab, "Semantic Classification Model for Twitter Dataset Using Wordnet" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 5, Issue 2, pp 5-9, February 2021. Article DOI https://doi.org/10.47001/IRJIET/2021.502002

****

