Sentiment Analysis on Twitter data with Semi-Supervised Doc2Vec

Metin BİLGİN[1], İzzet Fatih ŞENTÜRK[1]

[1]Faculty of Natural Sciences, Architecture and Engineering
Bursa Technical University
metin.bilgin@btu.edu.tr, izzet.senturk@btu.edu.tr

Abstract--**Twitter is one of the most popular microblog sites developed in recent years. Feelings are analysed on the messages shared on Twitter so that users ideas on the products and companies can be determined. Sentiment analysis helps companies to improve their products and services based on the feedback obtained from the users through Twitter. In this study, it was aimed to perform sentiment analysis on Turkish and English Twitter messages using Doc2Vec. The Doc2Vec algorithm was run on Positive, Negative and Neutral tagged data using the Semi-Supervised learning method and the results were recorded.**

Keywords--**Semi-Supervised Learning, Doc2Vec, Sentiment Analysis, Machine Learning, Natural Language Processing**

I. INTRODUCTION

Today, the use of the Internet is much higher compared to the past. Social platforms such as Facebook and Twitter have also been developed to enable people to quickly and easily share their views of any product or service. With the increase of social media content on the web, users can express their thoughts on social blogs such as personal blogs, Facebook and Twitter [1]. These shared messages constitute a rich and useful resource for research on social psychologists, marketing intelligence and intellectual mining [2, 3]. The increasing use of the internet has made it a great need to classify user comments as positive or negative.

It is of great importance to pre-process, process and classify large amount of data properly. It is possible to classify the data obtained through social sharing sites by running various machine learning algorithms. For machine learning algorithms to be able to classify data with high accuracy, the pre-processing phase must be correctly organized. The sentiment analysis of the data obtained from social networks is very difficult to perform because it contains misspelled words, abbreviations and words and phrases from the daily conversation jargon. For this reason, data must be filtered and processed with natural language processing methods [4].

Today, many studies on natural language processing have been performed on twitter data. Some examples of these studies are: [5] anticipating epidemics in May/December 2009 [5] analyzing medications and their unknown side effects from tweets between May 2009 and October 2010 [6], estimating social media dynamics and the change in human perception over time with a success of 85% using time series [7], and conducting perception analysis of over 70 million tweets that tourists have come across in a tourist destination [8]. In another study, the company's sales performance was attempted to be estimated by developing an opinion model based on the opinions written on blogs [9]. In [10], over three million tweets [11] measured the success of films that were shown by applying curve fitting to film critiques collected from different sources. In [12], estimation of stock market movements was tried by using Twitter data. In another study [13], positive and negative classifications were made based on comments shared from the IMDB movie evaluation platform. In [14], emotional orientations were determined with an accuracy of 74% from the comments made using an uncontrolled learning algorithm.

There are few studies on sentiment analysis using Twitter data in Turkish. There are studies which can determine whether the comments on a news website are positive or negative with a success rate of 85% [15], classification of movie reviews based on the positive and negative comments by using a dictionary based method [16], classification of tweets as positive, negative, or neutral using 2 and 3 gram model based on dictionary [17], attempting to find out if there is a relationship between the change in the stock market and the Twitter users' tweets about the economy [18], using two different data sets of Twitter and movie comments 75.2% to 85% success rate was achieved on the Twitter dataset using sentiment analysis using dictionary based machine learning methods, and achieved and 79.5% to 89% success rate on the movie reviews [19]. In [4], Turkish twitter messages are classified as positive, negative and neutral by using dictionary method and n-gram (2-3-4) model.

II. EXPERIMENTAL METHOD

A. Semi-Supervised Learning

It is one of the learning methods used in machine learning. The input data comprises large amount of

unlabeled data and small quantities of labeled data. This method is generally useful when the labeled data is low and the unlabeled data is readily available. Semi-Supervised learning is seen in Figure 1.

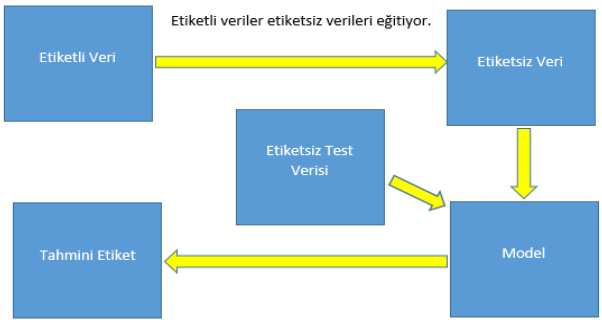

Figure 1: Semi-Supervised Learning

B. *Doc2Vec*

The Word2Vec algorithm is a method of constructing a vector by carrying words in a spatial manner. In Doc2Vec, the statements or paragraphs perform similar operations. In some sources Doc2Vec is referred to as Paragraph2Vec. This is the modified version of the Word2Vec algorithm [1,4,5]. The only change made according to the Word2Vec algorithm is the addition of a document ID, as can be seen in Figures 2 and 3. The goal of Word2Vec is to maximize the average probability in the training process.

$$\frac{1}{T}\sum_{i=k}^{T-k} \log p\ (w_i|w_{t-k}, \dots, w_{i+k}) \qquad (1)$$

Equation 1 gives the sequence of w_1, w_2, .., w_T training words.

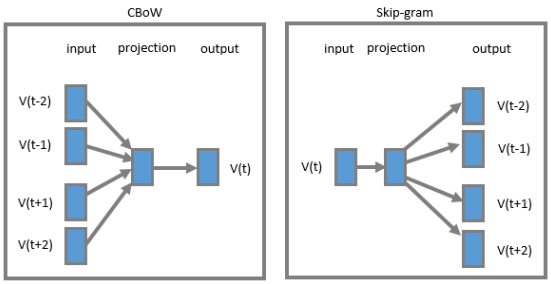

Figure 2. Word2Vec (CBOW and Skip-gram)

There are two different methods in the Word2Vec algorithm: Continuous bag of words (CBoW) and Skip-Gram (SG). These methods have been configured for Doc2Vec and have been translated into two methods: Distributed Memory (DM) and Distributed bag of words (DBoW).

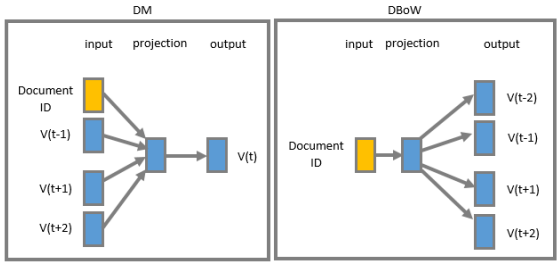

Figure 3: Doc2Vec (DM and DBoW)

C. *Lineer Regression*

In systems where variable quantities are present, the effect of some variables on others needs to be examined. The relationship between these variables is tried to be expressed by mathematical models. The equations which test whether there is a relation between two or more variables and express it linearly or curvilinearly are called regression model [20].

Linear regression is the method used to model the relationship between a dependent variable (y) and one or more independent variables. If there is a single argument, this is referred to as "Simple Linear Regression", if there are two or more independent variables, this is called "Multiple Linear Regression" [21].

In the given $\{y, x_1, \dots, x_i\}_{i=1}^{n}$ data set, it is assumed that the relationship between the dependent variable y and the independent variables vector x_i is linear. This linear relation is expressed by;

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon \qquad (2)$$

This is called the multiple regression equation. $\beta_0, \beta_1, .., \beta_i$ are known as the coefficients of the regression. In the case where there is only one argument, the relation is expressed by;

$$y = \beta_0 + \beta_1 X_1 \qquad (3)$$

This equation is called the simple linear regression equation.

D. *Data Sets*

In this study, two different data sets were used for Turkish and English.

Turkish dataset contains 2906 tweets for sentiment analysis [22]. The information regarding the data set is given in Table 1. The unlabeled training comprises 2281 sentences obtained through the Twitter API.

TABLE 1

TURKISH DATA SET

Class	The number of sentences
Labeled	
Positive	724
Negative	1270
Neutral	912
Toplam	2906
Unlabeled	
2281 Sentences	

English data set contains 1774 unlabelled and 58817 labeled data sets [23]. The information regarding the data set is shown in Table 2.

TABLE2

ENGLISH DATA SET

Class	Sample Size
Labeled	
Positive	301
Negative	1091
Neutral	382
Total	1774
Unlabeled	
58817 Sentences	

Before running Doc2Vec algorithms on data sets, some unnecessary data in the data must be cleaned. At this stage, called the preliminary phase, the data that is not a prelude to sentiment analysis should be deleted. This improves the success rate of the system and prevents unnecessary information from being dealt with, thus reducing the calculation cost of the system. Table 3 provides examples of cleaning.

TABLE 3

DELETED DATA AT THE PRELIMINARY PHASE

Deleted Character
Html Tags
Twitter User Name
Twitter Hashtags
Phone Number, Area Code, etc.
Mail Code
Regular Expression

The raw data obtained from the main servers of Twitter comes in Json format and the conversion for alphabetical characters specific to Turkish such as "ç" or "ş" directly affects the accuracy of the results to be obtained during the sentiment analysis phase. Example: In Json format, the expression "\ u00d6" in Turkish is the letter "Ö". The conversion process has been performed on the raw data.

III. EXPERIMENTAL RESULTS

Two different Doc2Vec algorithms have been employed for the two different data sets. The accuracy metric was used in the evaluation of the study results. Also, Zero Rule (R_0) value is calculated for Turkish and English data sets.

Accuracy metric, the most popular and simple method used to measure model performance, is the accuracy rate of the model. The number of correctly classified samples (TP + TN) is the ratio of the total sample counts (TP + TN + FP + FN). The error rate is 1 of this value. In other words, the number of misclassified samples (FP + FN) is the ratio of the total number of samples (TP + TN + FP + FN) [21].

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

The parameters used for the DM algorithm in the performed operation are shown in Table 4.

TABLE 4

PARAMETERS USED FOR DM

Parameter	Value
Size	400
Window	8
Min_count	1
Sample	1e-4
negative	5
workers	4
dm	1
dm_concat	1

For the DBoW algorithm, there is no dm_concat parameter, and the dm parameter has the value 0 (zero).

A software was developed using the Gensim library in the Python programming language for the work to be done. The software we developed is designed to be semi-consultant and able to train the system. Training and testing steps have been carried out by using DBoW and DM algorithms. In addition, the accuracy value for training and test set was calculated with the software and confusion matrices were created.

The results of the study for the Turkish data set are shown in Table 5. In addition, 2500 sentences for the test and the confusion matrix for the 406 labeled + 2281 untagged sentences used for the

training are shown in Table 6-7. For Turkish, R_0 is calculated equal to 0.43.

TABLE 5

RESULTS FOR TURKISH DATA SET

Approach	Number of Sentences						
Test	250	500	750	1000	1500	2000	2500
DM	0,416	0,448	0,4413	0,443	0,4306	0,4315	0,4328
DBoW	**0,44**	**0,452**	**0,4506**	**0,46**	**0,448**	**0,4405**	**0,4488**
Train							
DM	0,4397	0,4343	0,4359	0,4312	0,4459	0,4492	0,4433
DBoW	**0,4574**	**0,4546**	**0,4536**	**0,4449**	**0,4587**	**0,4569**	**0,4679**

TABLE 6

CONFUSION MATRIX FOR TURKISH TEST SET (THE NUMBER OF SENTENCES=2500)

DM	Negative	Neutral	Positive
Negative	1037	35	16
Neutral	728	18	35
Positive	577	24	25
DBoW	Negative	Neutral	Positive
Negative	1058	20	10
Neutral	723	30	28
Positive	575	22	34

TABLE 7

CONFUSION MATRIX FOR TURKISH TRAIN SET (THE NUMBER OF SENTENCES=406 LABELED + 2281 UNLABELED)

DM	Negative	Neutral	Positive
Negative	169	8	5
Neutral	123	5	3
Positive	84	3	6
DBoW	Negative	Neutral	Positive
Negative	177	4	1
Neutral	123	5	3
Positive	83	2	8

The results of the study for the English data set given in Table 2 are shown in Table 8. In addition, the confusion matrix for the 1500 sentences separated for the test and the 274 labeled + 58817 untagged sentences used for the training appear in Tables 9 and 10. For English, R_0 is calculated equal to 0.61.

TABLE8

RESULTS FOR ENGLISH DATA SET

Approach	The number of sentences					
Test	250	500	750	1000	1250	1500
DM	0.604	0.626	0.620	0.601	0,6128	0,6066
DBoW	**0.608**	**0.632**	**0.635**	**0.610**	**0,6216**	**0,6206**
Train	1524	1274	1024	774	524	274
DM	0.616	0.610	0.6113	0.631	0,6216	0,6131
DBoW	**0.635**	**0.6224**	**0.626**	**0.6421**	**0,6431**	**0,66**

TABLE9

CONFUSION MATRIX FOR ENGLISH TEST SET (THE NUMBER OF SENTENCES=1500)

DM	Negative	Neutral	Positive
Negative	862	50	10
Neutral	270	43	7
Positive	229	25	5
DBoW	Negative	Neutral	Positive
Negative	874	43	5
Neutral	265	49	6
Positive	225	25	8

TABLE10

CONFUSION MATRIX FOR ENGLISH TRAIN SET (THE NUMBER OF SENTENCES =274 LABELED+58817 UNLABELED)

DM	Negative	Neutral	Positive
Negative	151	10	8
Neutral	46	14	2
Positive	37	2	3
DBoW	Negative	Neutral	Positive
Negative	161	5	3
Neutral	46	16	0
Positive	37	2	4

English and Turkish datasets were run with semi-supervised learning and the results given in Tables 5 and 8 were obtained. The graphs for the obtained results are shown in Figures 4-7.

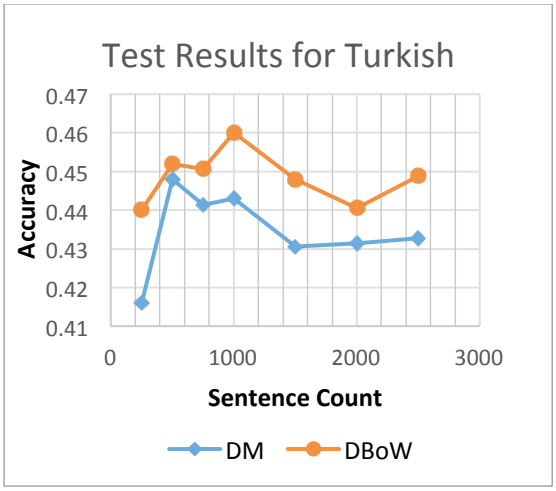

Figure 4 Test results for Turkish data set

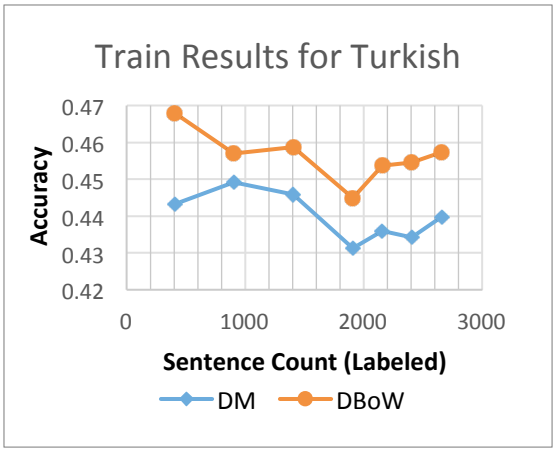

Figure 5 Training results for Turkish data set

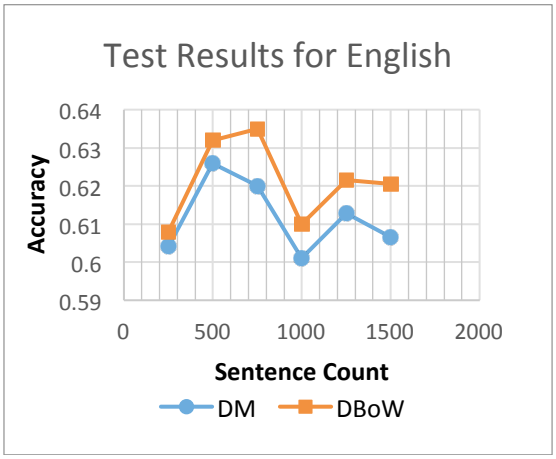

Figure 6 Test results for English data set

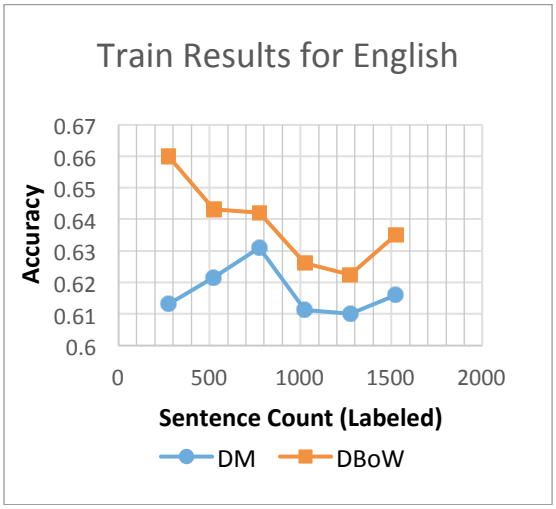

Figure 7 Training results for English data set

IV. CONCLUSIONS

In this study, Sentiment Analysis was performed using messages obtained from Twitter for two different languages. Some of the twitter messages obtained are primarily labeled as positive, negative, and neutral. Labeling was done for both Turkish and English datasets. Most of the data have not been included in the labeling process since half-advised learning should be done on the label-free data.

Our goal in this study is to train and test the system using a semi-supervised learning algorithm on twitter messages. Two different versions of the Doc2Vec algorithm, DM and DBoW, which were recently developed and continue to work on the model, have been used. The success of the methods in the test phase was measured following the modeling.

This study is important in the sense this work is one of the first studies employing Doc2Vec for sentiment analysis in Turkish. The study also shows that the DBOW method is more successful than DM. From this, it can be concluded that the operation algorithm of DBoW method is better than DM. The fact that the training and test results obtained for Turkish are lower than the English language indicates that the causal data set is smaller.

As a future work, it is our primary goal to enlarge the Turkish-labeled data set and to investigate the effects of the number of labeled data sets on the system's success. In addition to the semi-supervised learning, the program codes for the supervised learning will be updated and the system's success will be investigated. Another future plan is to investigate the use of DBOW and DM as a hybrid approach in emotional analysis and to obtain results.

REFERENCES

[1] Go, A., Huang, Lei, and Bhayani,R.. "Twitter sentiment analysis." Entropy 17 (2009).

[2] Bollen, J., Huina M., and Xiaojun Z., "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.

[3] Prabowo, R. and Thelwall, M., "Sentiment analysis: A combined approach." Journal of Informetrics 3.2 (2009): 143-157.

[4] Akgül, E.S., Ertano,C. ve Diri, B., "Twitter verileri ile duygu analizi.", Pamukkale University Journal of Engineering Sciences, 22(2), (2016): 106-110.

[5] Szomszor, M. N., Patty Kostkova, and Ed De Quincey. "# Swineflu: Twitter predicts swine flu outbreak in 2009.", 3rd International ICST Conference on Electronic Healthcare for the 21st Century (eHEALTH2010). 2012.

[6] Bian J, Topaloglu U, Yu F. "Towards large-scale Twitter mining for drug-related adverse events". International Workshop on Smart Health and Wellbeing (SHB'12), Maui, Hawaii, USA, 29 October-2 November 2012.

[7] Nguyen LE, Wu P, Chan W, Peng W, Zhang Y. "Predicting collective sentiment dynamics from time-series social media". Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '12), Beijing, China, 12 August 2012.

[8] Claster WB, Dinh H, Cooper M. "Naive bayes and unsupervised artificial neural nets for Cancun tourismsocial media data analysis". 2nd World Congress on Nature and Biologically Inspired Computing (NaBIC). Kitakyushu, Fukuoka, Japan, 15-17 December 2010.

[9] Liu Y, Huang X, An A, Yu X. “ARSA: A sentiment awaremodel for predicting sales performance using blogs”. 30th ACM SIGIR International Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, 23-27 July 2007.

[10] Asur S, Huberman BA. “Predicting the Future with Social Media”. IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), Toronto, ON, Canada, 31 August-3 September 2010.

[11] Joshi M, Das D, Gimpel K, Smith NA. “Movie reviews and revenues: an experiment in text regression”. Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Los Angeles, CA, USA, 1-6 June 2010.

[12] Bollen J, Mao H, Zeng X. “Twitter mood predicts the stock market”. Journal of Computational Science, 2(1), 1-8, 2011.

[13] Pang B, Lee L, Vaithyanathan S. “Thumbs up? Sentiment classification using machine learning techniques”. Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, USA, 6-7 July 2002.

[14] Turney PD. “Thumbs up or thumbs down? Semantic orientation Applied to unsupervised classification of reviews”. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 7- 12 July 2002.

[15] Eroğul U. Sentiment Analysis in Turkish. MSc Thesis, Middle East Technical University, Ankara, Turkey, 2009.

[16] Vural AG, Cambazoğlu BB, Şenkul P, Tokgöz ZO. “A frame work for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish”. 27th International Symposium on Computer and Information Sciences, Paris, France, 3-4 October 2012.

[17] Meral M, Diri B. “Twitter üzerinde duygu analizi”. IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Trabzon, Türkiye, 23-25 Nisan 2014.

[18] Şimşek M, Özdemir S. “Analysis of the relation between Turkish twitter messages and stock market index”. 6th International Conference on Application of Information and Communication Technologies (AICT), Tbilisi, Georgia, 17- 19 October 2012.

[19] Türkmenoğlu C, Tantuğ AC. “Sentiment analysis in Turkish media”. Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM ’14), Beijing, China, 21-26 June 2014.

[20] Öztürkcan M., "Regresyon Analizi", Maltepe Üniversitesi Yayınları Sayı:3, No:40 (2009).

[21] Kesim, M. "Real time measurement of micro changes in dinamic images", Msc. Thesis, Karadeniz Technical University, Trabzon, Turkey, 2015.

[22] Çetin, Mahmut, and M. Fatih Amasyalı. "Supervised and Traditional Term Weighting Methods for Sentiment Analysis.", SIU 2013, KKTC.

[23] Airline Twitter Sentiment, https://www.crowdflower.com/data-for-everyone/, Online: April 2017.