




Emotion Detection with n-stage Latent Dirichlet Allocation for Turkish Tweets

*¹Zekeriya Anıl Güven, ²Banu Diri, ³Tolgahan Cakaloğlu

¹Ege University, Faculty of Engineering, Computer Engineering Department, Izmir, Turkey, zekeriya.anil.guven@ege.edu.tr, 

²Yıldız Technical University, Faculty of Electrical-Electronics, Computer Engineering Department, Istanbul, Turkey, banu@ce.yildiz.edu.tr, 

³University of Arkansas, Computer Science and Computer Engineering Department, Arkansas, USA, txcakaloglu@ualr.edu.tr, 

Research Paper

Arrival Date: 12.09.2018

Accepted Date: 13.03.2019

Abstract

Understanding the reason behind the emotions placed in the social media plays a key role to learn mood characterization of any written texts that are not seen before. Knowing how to classify the mood characterization leads this technology to be useful in a variety of fields. The Latent Dirichlet Allocation (LDA), a topic modeling algorithm, was used to determine which emotions the tweets on Twitter had in the study. The dataset consists of 4000 tweets that are categorized into 5 different emotions that are anger, fear, happiness, sadness, and surprise. Zemberek, Snowball, and first 5 letters root extraction methods are used to create models. The generated models were tested by using the proposed n-stage LDA method. With the proposed method, we aimed to increase model's success rate by decreasing the number of words in the dictionary. Using the multi-stage LDA (2-stages:70.5%, 3-stages:76.375%) method, the success rate was increased compared to normal LDA (60.375%) for 5 class.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Natural Language Processing, Emotion Analysis

1. INTRODUCTION

Topic modeling determines the semantic structure of a text document. This method can organize and summarize the large-size text document [1]. Topic probabilities in a topic model provide an explicit indication that the document is understandable. This method can be successfully used in many areas like automatic document indexing, document classification, subject discovery and tendency analysis [2]. While the topics in the model are calculated as a probability distribution over the words; text documents are also calculated as a probability distribution over the topics [3]. Social medias have become more effective in many areas than communication. Users can share their thoughts and experiences about any subject through social medias such as Facebook. Those platforms can be also utilized for variety of purposes such as streaming news [4]. Emotion analysis is one of the most commonly studied social media research topics. Examples of positive and negative comments related to a shared topic such as identifying users' moods and ideas in the community are the examples of emotional analysis studies [5]. Additionally, emotion analysis is composed of 3 underlying topics that are document-based, sentence-based and feature-based. Classification of the document as positive or negative is called document-based. Sentence-based emotional analysis is performed for each of the sentences in the document. According to the emotional expressions that

characterize the features of the document, the classification as positive or negative for these features is defined as feature-based emotional analysis [6].

Throughout our literature survey, we mostly focused on articles that study text mining. Roberts et al. [7] studied emotion analysis in English based tweets where they benefited the LDA algorithm to extract the features. Çelikyılmaz et al. [8] applied the LDA model to a question answering system. The similarity score between question and candidate answers was performed by LDA. Çelikyılmaz et al. [9] studied a semantic process to understand speech. They also used secret N-gram clustering and semi-supervised LDA methods to learn the semantic structure of the speech comprehension system. The issue obtained by the developed LDA method has added an additional constraint to the learning model for the semantic structure. Paroubek et al. [10] completed a linguistic analysis of the collected document. After the feature extraction with the N-gram method, they created an emotion classifier that determines positive, negative and neutral emotions considering a whole document. They also used subject modeling methods to extract system features. Lin et al. [11] simultaneously developed product features and emotional expressions from cinema interpretations with the Joint Sentiment topic model that is based LDA method. Chatterjee et al. [12] implemented the Foreground Background Dice-LDA and

Reason Candidate and Background-LDA methods of the LDA for extracting topics from Twitter data; Senti Strength and semi-supervised Support Vector Machines are used for classifying emotions. Feuerriegel et al. [13] used the LDA method to remove the issues in the financial news where they also determined the impact of these issues on the German stock market. Mihalcea et al. [14] classified news headlines with their developed structure. This developed structure was used to find the link between emotions and words. Çoban et al. [5] proposed a method for analyzing the emotions in Turkish tweets. They tested the proposed system by using different feature extraction models and classifiers. They found that the success rate of emotion classification increased by 26%. Colace et al. [15] proposed a new emotion analysis approach to weighted word pairs obtained by using the LDA method in their work. In the proposed method, they aimed to determine a graphical model with a positive and negative attitude and a word based approach. Onan [16] assessed the predictive performances of the machine learning classifiers in emotion analysis by effectively representing Turkish tweets through LDA-based subject modeling.

There are many topic modeling methods for emotion detection. In this paper, we took the advantage of the LDA algorithm for detecting emotions detection stated in Turkish tweets. The algorithm has been developed to be n-stage for better emotion detection. Two and three stage developed with LDA method was compared with the tweets represented based on LDA in this paper.

In the second part of the paper, we will be detailing the dataset, preprocessing methodologies we experiment during our study. In the third part, we will present the empirical studies and their results on emotion analysis. Last but not least, the fourth part will criticize the results of the experiments.

2. MATERIALS AND METHODS

2.1. Latent Dirichlet Allocation Algorithm

LDA is a probabilistic topic modeling method that generates words and weights for a number of topics from a set of documents. In the LDA method, the text document is defined as the unified form of the subjects. On the basis of the method, the text has a probability distribution on the words, and the text documents have a probability distribution on the topics. Every subject has a distribution on the fixed word array [17]. The model aims to determine the basic structure of the subject with the words and weight values obtained from the observed dataset. The words in the documents are observations in the system.

LDA is an effective unsupervised learning method that is used to find topics in text documents. This method models each document as a mixture of each topic with a multi-term distribution over the words. The topic and topic-word distributions of the document learned by LDA define the best

topics for the documents. Also these distributions determines the most distinct words for each topic [18].

In LDA algorithm, all words in each document are randomly assigned topic. After random assignment, various statistics are extracted with this information. While local statistics are showing how many words are assigned to topics in each document, global statistics show how many words are assigned to each subject for the entire document. After the statistical information is obtained, the assignment of each word to each document is re-done. For this, the existing vocabulary is also updated.

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \quad (1)$$

The assignment of the words to the topics is calculated by looking at the relation of the document to topics (1). n_{ik} value indicates number of words assigned to the topic k in the new i. N_i is the total number of words in the document. The reason for subtracting 1 from value is to ignore the used word. α value gives the distribution of topics in documents. K value is also number of specified topics.

The number of topic K is determined by the coherence value that is the subject modeling criterion in the system. Coherence value measures the similarity of the words. Additionally, it provides the topic number to be selected. The k value that is the value of the highest outcome is selected as the number of topics among the coherence values calculated for the topic numbers mentioned above [19].

$$\frac{n_{word,k} + \beta}{\sum_{w \in V} n_{w,k} + V\beta} \quad (2)$$

In the method, secondly, it is calculated how much each word is related to the topics. The calculation yields the weight of each word in the topics. In Equation (2); $n_{word,k}$ indicates the number of times the current word is assigned to the k. topic in the entire document. The value of β gives the distribution of the words in the text. V is the size of the dictionary created from all the words in the dataset. By multiplying the values obtained by Equation (1) and (2), the probability of assigning the current word to the topic k is calculated. The values are recalculated throughout the number of all documents. The topic of the highest value is determined as the new topic of the word. The same operations are applied to every word of all the documents in the dataset to find the topics of the documents [20]. Updating the topics continues until the number of iterations specified in the system is satisfied. After having the topic distribution of the words, a document-term matrix is formed to extract the model of the system. By calculating the word weights with this matrix, weights of the words are obtained [21]. In the proposed method, the number of words in the dictionary of the entire document is reduced by the threshold value calculated using words and weights. This n-stage method aims to increasing success by weighing with less word at

every step. This n value may vary depending on the size of the dataset in the system.

The LDA structure is shown in Figure 1. Random variables are indicated by nodes. Possible connections between nodes are represented by using edges. In the Figure 1;

- α represents the topic distribution per document.
- β represents the word distribution per topic.
- Θ shows the topic distribution for a given document.
- z is an assigned topic for each word.
- w is an observed word.

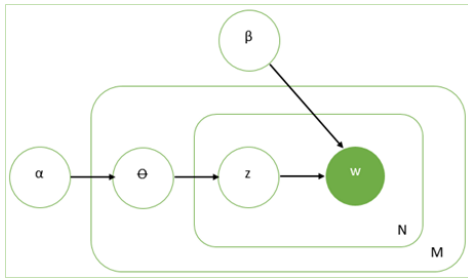


Figure 1. Latent Dirichlet Allocation structure [22]

As we inferred from the structure in Figure 1, the α and β parameters are sampled once as the system is initialized. The Θ parameter is sampled for each document in the system [22].

2.2. Dataset

Dataset is composed of Turkish tweets that are collected using the Twitter API. Instead of collecting all Turkish tweets, we applied a filtering that seeks at least one emotion expression (happiness, surprise, and etc) in a tweet. With this approach, we canceled out the redundant tweets. The dataset consists of five different emotions that are happiness, sadness, surprise, fear, and anger. We collected 800 tweets for each emotion. Two different datasets were created for use in training. These datasets are consist of 3 and 5 class labels respectively. Anger, fear and happiness labels are used for 3 class. Datasets contain 2400 tweets for 3 class and 4000 tweets for 5 class. 80% of the dataset was used for training and 20% for testing.

2.3. Preprocessing

Firstly, the punctuation marks are deleted in the tweets in the Turkish dataset. Then, all the tweets in the dataset are converted to lowercase. Since Turkish characters are faulty, non-English letters are translated into lowercase in the code. The very common stop words are removed from the tweets. Additionally, a list was created with meaningless words for emotions. The words in this list are removed from the tweets. Three different methods were used to find the roots of the words. Datasets were given for each of them with the names DB_z , DB_5 and DB_5 ;

- DB_z : The roots of the words were obtained by using the Zemberek library. This dataset was created of words including names, adjectives, verbs and reactions [23].

- DB_5 : The roots of the words were obtained by using the Snowball stemmer library. The first 5 letters were taken as root for those whose root length was longer than 7 characters [24]. The remaining words weren't changed.
- DB_5 : The first five letters of the words in dataset were taken as the root, and dataset was created.

Keşke hiç hayatımda olmasaydı dediğim çok insan var, onlar için üzülduğüme pişmanım

Zemberek (DB_z): keşke hiç hayat ol de çok insan var onlar için üzül pişman

Snowball (DB_5): keşke hiç hayat olmas dedik çok in var on iç üzülçük pişma

Get the first 5 letters as root (DB_5): keşke hiç hayat olmas dediğ çok insan var onlar için üzülđ pişma

Figure 2. An example tweet for root tools

Figure 2 shows updated version of the sample text with root finding tools. According to the figure, DB_z is promising since it outputs the results that can be interpreted easily as the morphological analysis of Turkish. Words in DB_5 and DB_5 have suffixes, thus some words have lost their meaning.

2.4. n-stage Latent Dirichlet Allocation

In order to increase system success, n-stage method is developed for LDA algorithm. The reason why we called the proposed system as n-stage is that the size of the dataset used in the system, or the amount of word associated with the topic in the tweets are dynamic. Figure 3 shows the steps of the method.

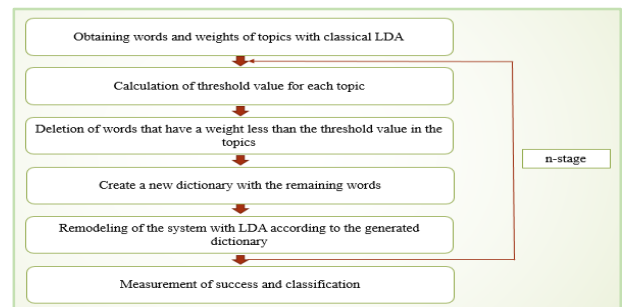


Figure 3. Structure of developed n-stage LDA method

We can explain the steps shown in Figure 3 as follows. Firstly, a threshold value is calculated for each topic from the words' weights in topics. Threshold value is obtained by proportioning total weight value of words belonging to topic to total number of words. This value is calculated for each topic. Then words with weight value less than threshold value for each topic are removed. Thus new dictionary is created with remaining words. Finally, system is re-modeled with LDA according to new dictionary.

Also, this method intended to reduce the number of words in the dictionary of the entire document. The reason why we are reducing the number of words in the dictionary is that words

with low weight in the model cause misclassification. Table 1 shows the number of words in the dictionary for the entire dataset. As the number of stages increases, it is clear that the number of words is decreasing in the dictionary.

Table 1. As the stage increases, the number of words in the dataset

Stage\Dataset	DB _Z	DB ₅	DB _S
1-stage	2208	4043	4291
2-stage	359	593	443
3-stage	309	168	149

As stage progresses, the number of words in dictionary decreases. Therefore, the weight values of remaining words change. Table 2 shows change in weight value of a sample word as stage progresses. As you can see, the weight value of word is increasing. Thus, class label of topic can be determined more easily.

Table 2. As the stage increases, the weight value change of the 'kork' word related to fear

Stage	Word Weight Value
1-stage	0.132
2-stage	0.343
3-stage	0.688

2.5. Programming Language and Platform

The study was developed with the Python¹ programming language. All processes such as preprocessing, read the datasets, development of the method were applied in the Python programming language.

Visual Studio program with plugin was used as the platform. In addition, Java platform was used for the Zemberek library just.

3. EXPERIMENTAL RESULTS

The number of studies done with Turkish emotional data is rather small. Datasets are often labeled as positive, negative, and neutral in related studies. Also, the datasets used in the literature aren't accessible. Therefore, developed model is compared only with classical LDA. LDA is an unsupervised method. n-stage LDA has been developed to increase the success of the system.

After applying pre-processing steps to datasets, coherence values are calculated for 3 and 5 class of generated DB_Z, DB_S and DB₅ datasets. 10 coherence values are calculated for each dataset. The value we use to train our system is topic number of the highest coherence value. For example,

coherence value and topic number found for DB_Z are given in Table 3.

Table 3. Number of topics with the highest value by emotion number (for DB_Z)

Emotion Number	Coherence Value	Topic Number
3	0.4998	9
5	0.484	20

The system is modeled with a specified topic number. Thus, a set of words and weight values are formed for each topic. Most appropriate class label is assigned according to words and weights. Table 4 shows examples of class labels assigned to topics. For example, topic 6 contains mostly sadness words. So, the label was assigned as sadness.

Table 4. Example of assignment of topic class labels

Topics	Words and Words' Weights	Labels
2	'0.132*"kork" + 0.052*"nefret" + 0.032*"korku" + 0.016*"hediye" + ...	Fear
4	'0.132*"scare" + 0.052*"hate" + 0.032*"fear" + 0.016*"gift" + ... '0.091*"sinir" + 0.076*"kafa" + 0.044*"irkil" + 0.042*"yiyecek" + ... '0.091*"anger" + 0.076*"flip" + 0.044*"blench" + 0.042*"out" + ...	Anger
6	'0.235*"mutsuz" + 0.113*"hüzün" + 0.031*"hasta" + 0.023*"tatlı" + ... '0.235*"unhappy" + 0.113*"sadness" + 0.031*"sick" + 0.023*"sweet" + ...	Sadness
14	'0.161*"yaşa" + 0.103*"günü" + 0.058*"doğum" + 0.046*"kutlu" + ... '0.161*"hooray" + 0.103*"day" + 0.058*"birth" + 0.046*"blessed" + ...	Happy
17	'0.201*"hayret" + 0.188*"şaşır" + 0.162*"şaşkın" + 0.051*"aaa" + ... '0.201*"wonder" + 0.188*"surprise" + 0.162*"confused" + 0.051*"aaa" + ...	Surprise

¹ <https://www.python.org/>

The success of the system constructed with LDA is reduced by increasing class number. Table 5 shows success of LDA for each dataset of 3 and 5 class.

Table 5. LDA's success compared to the root finding methods used

Emotion Number \ Dataset	DB ₅	DB _Z	DB _S
3	58	65.83	51
5	48.75	60.375	47

In order to increase the system's success, n-stage method is proposed for LDA algorithm. Word count in dictionary is reduced at each stage. Value of n in the method can vary according to size of the dataset. The process to be applied with increase of n is determined by threshold values of topics. Then a new dictionary is created with words that weigh more than the threshold value. The total word count in newly created dictionary is about 1/3 of previous one. Coherence values are re-calculated for 3 and 5 class, after topic numbers are re-determined. Table 6 shows results obtained when the system is re-modeled with two-stage LDA (2-LDA).

Table 6. The success of the system with the 2-LDA algorithm

Emotion Number \ Dataset	DB ₅	DB _Z	DB _S
3	68.5	80.83	74.375
5	67.375	70.5	56.875

System is re-modeled with a three-stage LDA (3-LDA) by selecting best resultant DB_Z dataset in two-stage LDA (2-LDA) method. For the third stage, new dictionary is created as before. Thus, word count in new dictionary is decreasing by half. Topic numbers for 3 and 5 class are determined by re-calculated coherence values. Table 7 shows system's 3-LDA model success and its comparison with 2-LDA model.

Table 7. 2-LDA with 3-LDA success comparison

Emotion Number \ Model	2-LDA (%)	3-LDA (%)
3	80.83	81.5
5	70.5	76.375

Table 7 shows the positive effect of developed n-stage LDA method. As n value increases, success rate increases linearly. For 3 class, success of 3-stage LDA method increased by approximately 1% compared to 2-stage LDA. For 5 classes, this increase is approximately 5%.

4. CONCLUSION AND DISCUSSION

In the study, it was used with LDA, which is a topic modeling algorithm, to detect emotion of the tweets in social media. Success of classical LDA method was compared with developed n-stage LDA. System success is increased between 10% and 15% in two-stage LDA method according to classical LDA method. When three-stage LDA method is applied, success rate is increased between 1% and 6% according to two-stage LDA.

Decreasing the word count in dictionary is the most important reason for increase in success. In this process, words with weight less than threshold value are removed from dictionary. Thus, as stage increases, weights of related words increases. This allows us to easily assign emotion labels to topics. For example, there are about 2700 words in classical LDA dictionary, while word count drops to 350 in two-stage LDA dictionary. Also, this count drops to 170 word in three-stage LDA. The n value in the method can be increased according to size of the dataset. It is reasonable to increase n stage, if word in tweets is less than emotional.

In our next topic modeling studies, we can use n-stage LDA algorithm to detect the music track, determine the effect on products of messages written in social media, find out which author wrote the text, find the correct answer in the question-answer systems.

In addition to developed method, a labeled dataset can be obtained from the word's weights in the topics, and we can also measure success rate in the classification algorithms.

REFERENCES

[1] D. M. Blei, "Probabilistic topic models", Communications of the ACM, vol. 55, no 4, pp. 77-84, April 2012.

[2] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models: a survey", Frontiers of Compute rScience in Chine, vol. 4, no 2, pp. 280-301, June 2010.

[3] M. Steyvers and T. Griffiths, "Probabilistic topic models", Handbook of latent semantic analysis, vol. 427, no 7, pp. 424-440, February 2007.

[4] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis", Mining text data, pp. 415-463, 2012.

[5] O. Coban, B. Ozyer, and G. T. Ozyer, "Sentiment analysis for Turkish Twitter feeds," 2015 23nd Signal Processing and Communications Applications Conference (SIU), May 2015.

[6] H. Türkmen, S. I. Omurca, E. Ekinici, "An Aspect Based Sentiment Analysis on Turkish Hotel Reviews", Girne American University Journal of Social and Applied Sciences, vol. 6, pp. 9-15, 2016.

[7] K. Roberts, M. Roach, J. Johnson, J. Guthrie, and S. Harabagi, "EmpaTweet: Annotating and Detecting Emotions on Twitter", In Proceedings of the 8th

International Conference on Language Resources and Evaluation (LREC), May 2012.

[8] A. Çelikyılmaz, G. Tur, and D. Tur, "LDA Based Similarity Modeling for Question Answering", Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, pp. 1-9, May 2010.

[9] G. Tur, A. Celikyilmaz, and D. Hakkani-Tur, "Latent semantic modeling for slot filling in conversational understanding," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 2013.

[10] P. Paroubek and A. Pak, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation, pp. 17-23, Malta, May 2010.

[11] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," Proceeding of the 18th ACM conference on Information and knowledge management - CIKM 09, pp. 375-384, Nov. 2009.

[12] R. Chatterjee and S. Agarwal, "Twitter Truths: Authenticating Analysis of Information Credibility", 2016 3rd International Conference on Computing for Sustainable Global Development, March 2016.

[13] A. Ratku, S. Feuerriegel, and D. Neumann, "Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation," SSRN Electronic Journal, pp. 1072-1081, Jan. 2014.

[14] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14," Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval 07, pp. 70-74, Jun. 2007.

[15] F. Colace, M. D. Santo, and L. Greco, "A Probabilistic Approach to Tweets Sentiment Classification," 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 37-42, Sep. 2013.

[16] A. Onan, "Türkçe Twitter Mesajlarında Gizli

Dirichlet Tahsisine Dayalı Duygu Analizi", Akademik Bilişim Konferansı, Feb. 2017.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993-1022, March 2003.

[18] L. Bolelli, Ş. Ertekin, and C. L. Giles, "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation," Lecture Notes in Computer Science Advances in Information Retrieval, pp. 776-780, Apr. 2009.

[19] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of Turkish Tweet Emotions by n-stage Latent Dirichlet Allocation", Electric Electronics, Computer Science, Biomedical Engineering's Meeting (EBBT), Apr. 2018.

[20] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of New Titles by Two Stage Latent Dirichlet Allocation", 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), Oct. 2018.

[21] J. Barber, "Latent Dirichlet Allocation (LDA) with Python," Human Activity Recognition Using Smartphones Data Set. [Online]. Available: https://studio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html. [Accessed: 12-Sep-2017].

[22] wikizero.net. [Online]. Available: <http://www.wikizero.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGF0ZW50X0RpcmljaGxl dF9hbGxvY2F0aW9u>. [Accessed: 20-Oct-2017].

[23] "Zemberek NLP," Zemberek NLP. [Online]. Available: <http://zembereknlp.blogspot.com/>. [Accessed: 05-Oct-2017].

[24] "Download," Snowball. [Online]. Available: <http://snowball.tartarus.org/download.html>. [Accessed: 16-Nov-2017].