

# Emoji Prediction From Text as a Sentiment Analysis Method

Beyda Güler, Şeymanur Korkmaz  
Bilgisayar Mühendisliği Bölümü  
Yıldız Teknik Üniversitesi, 34220 İstanbul, Türkiye  
{beyda.guler, seymanur.korkmaz}@yildiz.edu.tr

**Özetçe** —Sosyal mecralar aracılığıyla her saniye milyonlarca yeni veri girişi söz konusu olmaktadır. Bu durum, verilerin işlenmesi ve yorumlanması ihtiyacını doğurmuştur. Söz konusu verinin metin veya ses olması halinde devreye yapay zekanın bir alt kolu olan doğal dil işleme girmektedir. Yazılı metinler duygudan kısmen yoksundur. Emojiler, metinlerin duygu iletimine katkıda bulunan unsurlardır. Bu çalışmamızda salt metin içeren tweet'lere çeşitli makine öğrenmesi ve derin öğrenme algoritmaları kullanarak emoji tahmini yapılmaktadır.

**Anahtar Kelimeler**—doğal dil işleme, makine öğrenmesi, derin öğrenme, duygu analizi, emoji tahmini.

**Abstract**—There are millions of new data entries every second through social media. This situation has led to the need for processing and interpretation of data. If the data is text or voice, natural language processing, a sub-branch of artificial intelligence, comes into play. Written texts are partially devoid of emotion. Emojis are elements that contribute to the emotional transmission of texts. In this study, we predict emoji for text-only tweets using various machine learning and deep learning algorithms.

**Keywords**—natural language processing, machine learning, deep learning, sentiment analysis, emoji prediction.

## I. INTRODUCTION

Natural language processing is a branch of artificial intelligence that enables computers to comprehend, generate and manipulate human language. There are two different types of language for a computer. One is machine languages, i.e. programming languages, and the other is natural languages. Natural language processing, a branch of artificial intelligence, aims to analyse and understand or reproduce the structure of natural languages. Thanks to natural language processing, computers understand human language and give output in a way that humans can understand.

Due to the variety of usage areas and the convenience it provides, natural language processing is encountered in many areas of life. Today, at the last point of digitalisation, the number of electronic devices has surpassed the human population. This situation has led people to produce content for digital media.

Users not only produce content but also share all kinds of phenomena they observe in current life on the internet. This keeps the flow of information on the internet constantly alive. Data flow to the internet from many different platforms such as Twitter, Instagram, Twitch. When

we analyse Twitter, which is a good raw material for natural language processing studies, we see that 347.200 tweets are shared per minute. Natural language processing methods are used to make sense of this data, which is increasing exponentially every second.

Only 7-15% of the message desired to be given in communication consists of words. Apart from this, it is calculated that tone of voice and concept are 30%, gestures, mimics and facial expressions are 55% [1]. For this reason, an important part of the message to be given consists of emotional elements. Emojis are used to increase the emotion in texts. In order to fully comprehend the meanings of these texts, emotion analysis is performed.

In this study, it is aimed to perform sentiment analysis from Twitter data. Sentiment analysis studies generally focus on neutral, negative and positive meanings, but in this study, classification is applied over 20 different classes. [2]

The training set of Semeval Task-2 consists of 500.000 English and 100.000 Spanish tweets and most frequent 20 emojis for English and 19 emojis for Spanish data set. In this study the English data set is used.

0	1	2	3	4	5	6	7	8	9
❤️	😄	😁	❤️	🔥	😊	😎	🌟	💙	😏
10	11	12	13	14	15	16	17	18	19
📷	us	☀️	💜	😊	🎉	😁	🎄	📷	😏

Figure 1 Emoji Table

## II. METHOD

The data set initially contains 500.000 tweets and it is reduced to 346.748, partly because some of them were not available by the time we crawled them and partly because of the normalization process, which involved deleting empty lines and duplicated data. Emoji distribution in the data set is seen in Figure 2.

Some normalization steps have been applied to reduce the noise of data and to make the slang language used in social media more uniform. '@' and '#', numerical data,

'RT' expression, punctuation marks and HTML characters removed from tweets as the first normalization step. Then words with the same meaning but written in different forms were identified and reduced to a single form.

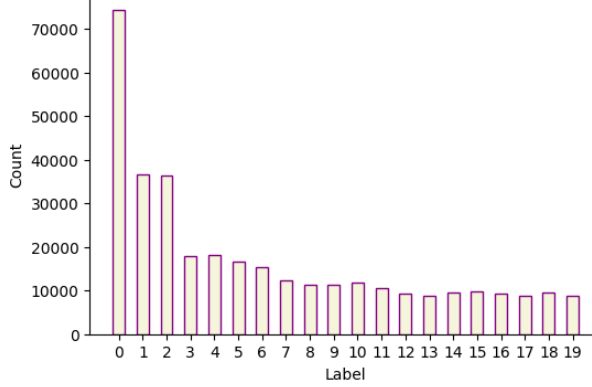


Figure 2 Emoji Distributions

10% of the dataset is reserved for testing and 90% for training. For the validation dataset, which is used to evaluate the performance of the model during training, 20% of the 90% data is reserved as validation data.

Within the scope of our project, traditional and modern modelling has been used. These are: LSTM (Long Short-Term Memory), Bidirectional LSTM and BERT (Bidirectional Encoder Representations from Transformers) models. In LSTM and BLSTM models, three different embedding models were used: Glove, Word2Vec and Keras.

Although in BERT, traditional word-level embedding models, such as Word2Vec or GloVe, are not used. Instead of these models a combination of token embedding, segment embedding and position embedding models is used. Two different oversampling methods, ROS and SMOTE technique was used to obtain various results.

This article analyses the two highest performing models and the application of two-stage classification to the highest performing versions of these models. These are: Bidirectional LSTM, BERT, Two Stage Classification applied on BLSTM and Two Stage Classification applied on BERT.

#### A. Bidirectional LSTM

Bidirectional LSTM networks function by presenting each training sequence forward and backward to two independent LSTM networks, both of which are coupled to the same output layer.

Rather than encoding the sequence in the forward direction only, BLSTM encodes it in the backward direction as well and concatenate the results from both forward and backward LSTM at each time step. The encoded representation of each word understands the words before and after the specific word [3].

Basic architecture of Bi-LSTM is shown in Figure 3.

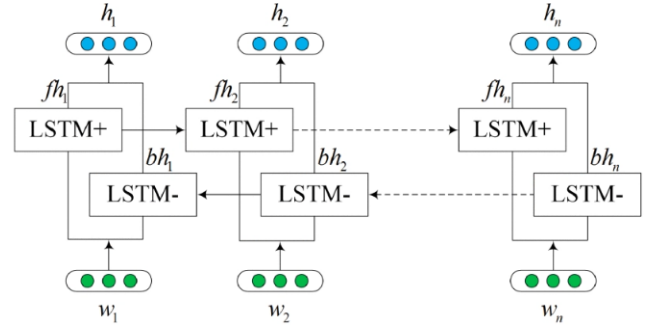


Figure 3 BLSTM Architecture [3]

When we look at the Table-1, we see that the highest success rate in BLSTM models is 0.29 when no oversampling is performed and using the GloVe embedding model.

Table 1 F-1 Score Results for BLSTM Models

	Word2Vec	GloVe	Keras
<b>Oversampling - ROS</b>	0.26	0.21	0.19
<b>Oversampling - SMOTE</b>	0.23	0.21	0.18
<b>Without Oversampling</b>	0.27	0.29	0.19

#### B. BERT

BERT is Google's neural network-based technique for natural language processing pre-training. Through this technique, Google better analyses users' search queries and serves users with more accurate results.

BERT is based on transformer architecture and unlike other models, it evaluates the sentence both from left to right and from right to left. In this way, the meaning and the relationships between words are better determined [4]. It can learn the context of a word in the text by looking at the words that come before and after it.

In addition to being bidirectional, BERT is trained with two techniques called Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). It is quite costly to train the transformer model with such a large data set using MLM and NSP techniques.

Therefore, pre-trained models are used to solve new problems by using a technique called fine-tuning.

Table 2 F-1 Score Results for BERT Models

	Combined Embedding
<b>Oversampling - ROS</b>	0.26
<b>Without Oversampling</b>	0.25

In the model results, it is observed that emoji with 0 label is very dominant in the predicted emojis. This problem should be solved because it affects the overall results in a negative way. For this reason, Two Stage Classification is developed.

### C. Two Stage Classification

Among the models applied, the ones with the highest success are identified. These models are BLSTM without oversampling with the GloVe embedding model (0.29) and BERT with Random Oversampling (0.26). These models are chosen to implement two-stage classification. However, in the BERT model with Random Oversampling, there is no obvious clustering for class 0. Therefore, instead of this model, two-stage classification is tested on the BERT model without oversampling (0.25).

In the first step, it is decided whether the input sentence belongs to emoji-0 or any of the other 19 emojis. In the second step, if the sentence belongs to one of the other 19 emojis, it is decided which emoji it belongs to.

**Table 3** F-1 Score Results for Two Stage Classification

	Two Stage Classification
BLSTM (GloVe&Without OS)	0.27
BERT (Without OS)	0.27

### III. RESULTS & DISCUSSIONS

In this section we will discuss the final results and the factors that reduced success.

As we already mentioned, oversampling is needed for this data set because the classes are unbalanced. ROS and SMOTE were used for oversampling. When the F-1 scores for each class are analysed, it is seen that SMOTE is more unsuccessful and does not prevent unbalanced distribution.

Using two-stage versions of the bert and blstm models, F-1 scores are increased to 27%. When the previous studies in the literature are examined, it is seen that the score remains at these levels. There are several reasons for this situation.








Firstly, normalization is an important step in NLP and when we look at Semeval Task-2 data set, it is seen that raw data is really noisy. Because the texts we have are tweets, which is real world data. Although this noise is reduced by applying normalization steps, there are unpredictable written expressions and misspellings in the data set. This situation affects the success to a great extent.

The second reason is that the emojis in the given data set have closer meanings to each other. As seen in the Figure 1; 0 (Red Heart), 1 (Smiling Face with Heart-Eyes), 3 (Two Hearts), 8 (Purple Heart) and 13 (Blue Heart) are similar emojis that are interchangeable in everyday speech. Therefore, it was observed that the models suggest similar emojis to each other. There is an example of this situation in Figure 4.

Tweet	True Emoji	Predicted Emoji
anna hopiefirst day nd grade kindestlittlegirl		

**Figure 4**

In order to avoid this situation, similar classes are combined and F-1 scores are recalculated for each class and the overall F-1 score increased to 34%. You can see a list of these similar classes in Figure 5. This improvement has increased the score, but still not at the desired rate. It is clear from this, the confusion of similar emojis is not the only factor decreases the success.

0	1	2	3	4	5	6	7	8	9	10	11	12	13
													

**Figure 5**

When the confusion matrix is analysed, it is seen that the 17th emoji (Christmas Tree) has the highest scores compared to the others. The reason for this is that tweets containing 'christmas' word are generally use this emoji and prediction becomes much more accurate than other emojis. It is also seen that emoji 0 generally has the highest success. The reason for this is, this emoji is the most common emoji in the data set, as can be seen from the data distribution plot.

In the two-stage classification models, we see that from Table 4 , the classes with 0.00 results are almost nonexistent. Likewise, the difference in success between the classes is also lower than the other models. It can also be said that the two-stage classification application generally increases the success of the models.

**Table 4** F-1 Score Comparison for Two Stage Classification Models

	LSTM	BLSTM	BERT
0	0.29	0.31	0.41
1	0.29	0.31	0.32
2	0.45	0.43	0.49
3	0.19	0.19	0.19
4	0.42	0.47	0.51
5	0.14	0.14	0.15
6	0.14	0.14	0.15
7	0.22	0.25	0.26
8	0.10	0.13	0.16
9	0.15	0.14	0.17
10	0.23	0.24	0.29
11	0.49	0.50	0.58
12	0.40	0.39	0.42
13	0.05	0.07	0.08
14	0.03	0.02	0.10
15	0.24	0.20	0.23
16	0.02	0.01	0.04
17	0.61	0.64	0.64
18	0.06	0.07	0.18
19	0.00	0.00	0.02
macro f-1	0.23	0.27	0.27

### A. Evaluation

The overall results show that the GloVe embedding model as the embedding model, the case without oversampling, and the BLSTM and BERT models among the models have higher success. Among the oversampling methods, ROS is more successful than SMOTE, but in some cases model without oversampling is more successful.

### REFERENCES

- [1] N. ÇALIŞKAN and R. Yeşil, “Eğitim sürecinde öğretmenin beden dili,” *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, vol. 6, no. 1, pp. 199–207, 2005.
- [2] Ç. Çöltekin and T. Rama, “Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction,” in *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 34–38. [Online]. Available: <https://aclanthology.org/S18-1004>
- [3] [Online]. Available: <https://www.codingninjas.com/codestudio/library/bidirectional-lstm>
- [4] [Online]. Available: <https://medium.com/@toprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611>