Detection of the Trends of Football Fans in Turkey by Machine Learning Algorithm Using Twitter Data

Erol Kına

Abstract—Twitter is a social media platform whose user base is growing day by day and where people can express their feelings in writing. People share their opinions on a current topic with texts called tweets, which allow them to be informed about the views they hold. In this study, the tweets on the social media platform Twitter were used to infer the football teams supported by users through classification with machine learning algorithms. A success rate of around 96% was achieved in the classification results.

Index Terms— Twitter, Sentiment Analysis, Machine Learning, Football fans, NLP, KNN, Classification

1 Introduction

THE use of social media platforms by users for reasons such as satisfying their daily needs, passing time, and research increases the rate of Internet use. According to the 2021 report published by TUIK, the percentage of internet users in Turkey is 82.6%. Based on the tweets shared by users on Twitter, data mining methods are used to classify their views.

Since the results of sports clubs in the matches and the team that will finish the season as champions in our country are always a matter of curiosity, many tweets about football teams are sent both during the matches and during the season. In this study, the sports clubs that users support were identified. In classifying the teams supported by Twitter users, a dataset was first created, the data was preprocessed, feature extraction was performed, the classification was applied, and the results were obtained. Tweets shared by users on Twitter were used to create the dataset. Tweets from the year 2022 were used with the help of the Twitter API. In this study, a total of 160000 tweets from 4000 users supporting 4 different sports clubs were used to create the dataset.

Users often use abbreviations because a certain number of characters must be used in posts on Twitter. For these reasons, the data obtained is noisy in its structure. For this purpose, natural language processing (DDI) methods are used when cleaning the data [2]. By using DDI methods, unnecessary characters and meaningless structures are removed from the text and noisy data is standardized. In this study, the data was cleaned using the zemberek library. With the zemberek library, the corpus of the word (term) is obtained by removing inflectional suffixes. The list of terms was defined by extracting the attributes of the data from 4 sports clubs using the term frequency (TF) method. The most frequently repeated terms were weighted and a total of 400 terms were selected, 100 for each sports club. The frequency ratio of these terms was determined using the data from the sports clubs. This procedure makes the data available for analysis with machine learning algorithms [3]. A training set was created to learn how to classify data for analysis. In this study, using classification algorithms, the values of Kappa statistics were used to compare the success rates [4]. While Naive Bayes (NB), Logistic Regression (LR), Random Forest Algorithm (RF) and K-Nearest Neighbor (KNN) algorithms were used in the study, it was found that the most successful results were obtained using the KNN algorithm with 96%.

In determining the results, user-supported teams of Galatasaray (GS), Fenerbahçe (FB), Beşiktaş (BJK), Trabzonspor (TS) in Super League were taken as reference. If users did not share any messages related to the teams they supported or if no opinion could be determined based on the tweets they sent, our result value was accepted as the exclusion of 4 sports clubs.

The sections of this study are organized as follows. The second section deals with the acquisition of the dataset and the third section deals with the preprocessing steps for the dataset. The fourth section describes the feature extraction from the processed data and the fifth section presents the experimental results. In the sixth section, the results of the study are mentioned.

2 ACQUISITION OF THE DATASET

When the dataset is created, users' tweets consist of positive opinions about the teams they support with Twitter API support. The users' positive comments about the matches, transfers and results make up the dataset. In creating the dataset, a total of 160000 tweets from 4000 users supporting 4 different sports clubs were used. The tweets of GS, BJK, FB and TS fans who have the largest fan base in Super League were used. A total of 40000 tweets from 1000 different users supporting the football club GS form the GS dataset. A total of 40000 tweets from 1000 different users supporting the football club BJK form the BJK dataset. A total of 40000 tweets from 1000 differ-

- *Erol Kına is currently pursuing Lecturer in Özalp Vozational School in Van Yüzüncü Yıl University, Turkey. E-mail: erolkina@yyu.edu.tr*

ent users supporting the football club FB form the FB dataset. A total of 40000 tweets from 1000 different users supporting TS football club form the TS dataset. Balanced datasets are preferred because they give more successful results. [4].

3 DATASET PRE-PROCESSING STEPS

Due to the use of abbreviations, emojis, special characters, numbers, and conjunctions in the comments shared by users on social media platforms, the data forming the dataset is distorted and a standard cannot be achieved [5]. NLP methods are used to overcome this difficulty. NLP methods are used to convert the corruption in the data into a standard form by removing unnecessary characters and noise. In this study, zemberek library among Turkish natural language processing libraries was used. The words are separated into their roots and arranged accordingly for feature extraction.

4 ATTRIBUTE INFERENCE

After the preprocessing steps of the defined data in the dataset, a list of sports terms was created and weighted by frequency [6]. For GS term list presented in Table 1 on GS dataset, for BJK term list presented in Table 2 on BJK dataset, for FB term list presented in Table 3 on FB dataset, for TS term list presented in Table 4 on TS dataset.

TABLE 1
GS WORD LIST

Order	Word	Frequency
1	Fatih	710
2	Terim	685
3	Elmas	609
4	Yönetim	565
5	İstifa	544
…	…	…
100	Kaleci	23

TABLE 2
BJK WORD LIST

Order	Word	Frequency
1	Sergen	962
2	Şampiyon	950
3	Yalçın	470
4	Teknik	460
5	Transfer	440
…	…	…
100	Güven	35

TABLE 3
FB WORD LIST

Order	Word	Frequency
1	Ali	1230
2	Koç	920
3	Yönetim	557
4	İstifa	515
5	Taraftar	503
…	…	…
100	Bulut	21

TABLO 4
TS WORD LIST

Order	Word	Frequency
1	Şampiyon	1250
2	Trabzon	940
3	Avcı	560
4	Sene	520
5	Abdullah	516
…	…	…
100	Ceza	34

For each sports club, the terms with the 100 highest frequency values were selected. Then, the term frequency method was used for feature extraction. The frequency of repetition of words in tweets was determined using the term frequency method. The frequency of repetition was calculated separately for each sports club. The obtained value is obtained by dividing the terms in the user's tweet by the total frequency count (TFS). GSTFS for GS, BJKTFS for BJK, FBTFS for FB, TSTFS for TS are calculated. The calculations are shown in equation 1-4.

$$GSTFS = \sum_{1}^{100} \left(\frac{Frequency\ (k)}{TFS} \right) \tag{1}$$

$$BJKTFS = \sum_{1}^{100} \left(\frac{Frequency\ (k)}{TFS} \right) \tag{2}$$

$$FBTFS = \sum_{1}^{100} \left(\frac{Frequency\ (k)}{TFS} \right) \tag{3}$$

$$TSTFS = \sum_{1}^{100} \left(\frac{Frequency\ (k)}{TFS} \right) \tag{4}$$

Immediately after feature extraction, a training set is created using machine learning algorithms. The qualifications and class labels of the four sports clubs are used in the creation of the training set. The reason for using class labels is to determine the class to which the data belongs. The data of a total of 80 users, 20 for each sports club, were used in the training set. The classification process for the data to be tested can be done after this stage. The training dataeset is shown in Table 5.

TABLE 5
TRAINING DATASET

	GSTFS	BJKTFS	FBTFS	TSTFS	Class
1	0.052568	0.078541	0.056151	0.138547	GS
2	0.2129635	0.1274585	0.152415	0.132584	BJK
3	0.145416	0.076874	0.199505	0.057395	FB
4	0.191325	0.102354	0.129741	0.050135	TS
5	0.154120	0.124712	0.135415	0.112547	GS
…	…	…	…	…	…
80	0.081212	0.046271	0.120085	0.062541	TS

5 CLASSIFICATION AND EXPERIMENTAL RESULTS

Classification algorithms are used to classify users' opinions about their teams. In this study, the success rates of the classification algorithms were analyzed using the values of the Kappa statistic [7]. The success rates were compared with the algorithms NB, LR, RF, and KNN. When comparing the success rates, the values of the Kappa statistics were used. A kappa value gives a result in the range of -1 to 1. If the kappa value is equal to 1, the reliability is complete. Reliability is in question when the kappa value is in the range of 0 and 1. If Kappa is less than 0, there is no reliability [7]. The calculation of the Kappa value is shown in equation 5.

$$K = \left(\frac{P_0 - P_c}{1 - P_c}\right) \qquad (5)$$

Po represents the accepted rate and Pc represents the accepted rate. The KNN classification algorithm is a distance-based computation algorithm. It uses feature similarity to estimate the values of new data points [8]. The computation process after Euclidean distance estimation is shown in Equation 6.

$$distance\ (x, y) = \sqrt{\sum_{1}^{i} (X_i - y_i)^2} \qquad (6)$$

When comparing the algorithms NB, LR, RF and KNN, the most successful result was achieved with the KNN algorithm. When examining the test results presented in Table 6, the KNN algorithm is the most successful algorithm according to the kappa statistics test for the model created. The kappa statistical value closest to 1 belongs to the KNN classification algorithm.

TABLE 6
TEST DATASET RESULTS (BALANCED DISTRIBUTION)

Algorithm	Percentage of correctly classified data	Percentage of incorrectly classified data	Kappa Test Result
NB	77%	23%	0.7385
LR	80%	20%	0.7567
RF	86%	14%	0.8927
KNN	96%	4%	0.9114

To compare the success rate for balanced and unevenly distributed data sets, Table 7 presents the success rates for unbalanced data sets. In creating the unbalanced dataset, we used the opinions of 3000 GS users, 1000 BJK users, 1090 FB users, and 1250 TS users.

TABLE 7
TEST DATASET RESULTS (UNBALANCED DISTRIBUTION)

Algorithm	Percentage of correctly classified data	Percentage of incorrectly classified data	Kappa Test Result
NB	56%	44%	0.5217
LR	61%	39%	0.5841
RF	67%	33%	0.6521
KNN	72%	28%	0.6994

Looking at Table 6 and Table 7, it is clear that the balance of the data set directly affects the success rate. The results were obtained by classifying the football clubs that support the users. If the users did not share any sports news or if no opinion could be detected, our result value was accepted as the exclusion of 4 football clubs.

6 CONCLUSION

Twitter allows us to obtain information by accessing data through the social media platform. In this study, the tweets in which users share the trends of football fans are examined. In creating the dataset, 4000 users were selected from among the fans of 4 different sports clubs. A dataset of 160000 tweets belonging to these users was created. Using data mining and DDI methods, features were extracted and a training set was created. The test data was tested using the Kappa statistics test. The classification algorithms KNN, NB, LR and RF were used in the testing phase. According to the test results, the KNN classification algorithm was the most successful classification algorithm with 96% correct classification and a kappa value of 0.9114. When the results of the uniformly distributed test data set are compared with the results of the non-uniformly distributed test data set, it can be seen that the uniform distribution of the data set is an important factor for the success rate.

REFERENCES

[1] [1] Türkiye İstatistik Kurumu Hanehalkı Bilişim Teknolojileri Kullanım Araştırması, https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-(BT)-Kullanim-Arastirmasi-2021-37437 . 2021.

[2] [2] E. Adalı, Doğal Dil İşleme. *Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 5(2), 2012.

[3] [3] Ö. Çoban, G. T. Özyer, Twitter duygu analizinde terim ağırlıklandırma yönteminin etkisi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24(2), 283-291, 2018.

[4] [4] N. Hatice, S. S. Akın, "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması," XIX. Türkiye'de İnternet Konferansı, May 2014.

[5] [5] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, F. Herrera, Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), 1-22, 2016.

[6] [6] E. E. Eryılmaz, D. Ö. Şahin, E. Kılıç, Türkçe İstenmeyen E-postaların Farklı Öznitelik Seçim Yöntemleri Kullanılarak Makine Öğrenmesi Algoritmaları ile Tespit Edilmesi. *Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 13(2), 57-77, 2020.

[7] [7] Kiliç, S, Kappa test. Psychiatry and Behavioral Sciences, 5(3), 142, 2015.