The Perception of Roma in Turkish Society:

A Sentiment Analysis of Tweets

By
Tunay Tokmak

Submitted to
Central European University Romani Studies Program

In partial fulfilment for International Interdisciplinary Romani Studies Postgraduate Specialization Program

Supervisor: Pál Galicza
Budapest, Hungary
2022

Table of Contents

The Perception of Roma in Turkish Society: A Sentiment Analysis of Tweets

Abstract

The Roma constitute a significant minority group in Turkey that is marginalized and stigmatized by mainstream society. The increase in social media usage results in a vast amount of data that can be used to derive public opinion on a particular issue and make data-driven decisions. In this context, sentiment analysis is implemented as a natural language processing technique to understand the perception of the Roma in Turkish society based on tweets posted with the *çingene* keyword, the equivalence of 'gypsy' in Turkish. The logistic regression model was used as the classifier while three different text vectorization techniques were applied. The model's precision differs based on the text vectorization techniques. It is 84.83 per cent for positive/negative word frequencies and the bag of words when it is 79.31 for term frequency-inverse document frequency. The analysis of collected tweets demonstrates that the Turkish society has a negative perception of the Roma.

Keywords: Gypsy, sentiment analysis, machine learning, logistic regression, Python

Chapter 1: Introduction

Romani people form a significant minority group in Turkey, whose population consists of multiple ethnic and racial groups. The Roma population in Turkey is estimated at 500,000 based on unofficial data, which is approximately 0.57 percent of the total population. (Alp 2016, 153) They differ significantly from mainstream society due to their distinct culture and are exposed to marginalization, racism, and stigmatization. The prejudice and hate speech toward the Roma manifest themselves in the Turkish language and legends. The Romani people are typically depicted as inferior, immoral, and disorderly. Phrases such as a 'gypsy tent' (*çingene çadırı*), 'gypsy wedding' (*çingene düğünü*), and 'gypsy pilav' (*çingene pilavı*) are used by Turkish society to emphasize disorder.

Also, the Roma are portrayed as shameless, thieves, officious, and barefaced in numerous Turkish novels (Kolukirik 2007, 34). The Roma are seen as half of a nation *(buçuk millet*) that does not have a moral code, identity, and religious belief. 'Gypsy faith' *(çingene imanı*) is an expression in the Turkish language to make fun of a person's faith. Additionally, Roma women are belittled in the language. Proverbs such as "A Gypsy woman does not make a wife" (*Çingeneden karı olmaz.*), and "Everybody chews gum, but the gypsy girl enjoys it" (*Herkes sakız çiğner ama çingene kızı tadını çıkarır.)* demonstrate the objectification of Roma women and sexism towards them in Turkey.

With the increase in social media usage, individuals can declare their thoughts and feelings on social forums. However, this convenience may result in the reproduction and dissemination of hate speech significantly fast. For instance, in 2015, a thread was opened on Sour Dictionary, a popular social forum in Turkey, asking what the word "gypsy" evoked in members' minds. As a result of the scrutinization of entries, nearly 69 percent were found to convey negative sentiments, while merely 28 percent were interpreted as positive entries (Alp 2016, 160).

Twitter is a major micro-blogging social media platform with over 200 million users worldwide (Twitter 2021). 8.31 percent of Turkey's population are Twitter users in 2022. People use it to produce knowledge, disseminate the news, generate discussions, and declare opinions. Therefore, Twitter is a proper platform to understand public opinion on a particular issue. In the context of the Roma, by understanding the motives behind the negative sentiments of users, policies regarding the Romani people can be made consciously. Parallel to this, the research aims to investigate the perception of the Roma in Turkish society by conducting sentiment analysis using a machine learning approach based on the analysis of tweets posted between 17th of April 2022 and 24th of May. 2022.

The study is organized as follows: in the first part, the literature related to the sentiment analysis of Turkish tweets by the machine learning approach is summarized. The second part examines the methodology of this study. Finally, the results and the limitations of the study are presented.

Chapter 2: Literature Review

2.1 Sentiment Analysis on Tweets

Improvements in artificial intelligence technology make it possible to gain insights from big data. Sentiment analysis is a widely used big data analytics technique to understand the opinions of the masses on a particular subject on online social media platforms. Twitter has become a popular channel for analysts to employ sentiment analysis due to the availability of Twitter's Streaming Application Programming Interface (API). Twitter's API is freely available, and grants users access permission to subsets of public tweets (Nyugen et al. 2018, 309).

Performing sentiment analysis on Twitter posts supports the strategic decision-making process regarding marketing policies, estimation of financial ratios, customer loss analysis, identifying opportunity and threat factors in the sector, and predicting competitors' activities. (Onan 2017, 4) Further, researchers utilize sentiment analysis to forecast election results and survey public opinion (Tokçaer 2021, 1515).

Twitter Sentiment Analysis focuses on the sentiment classification of individual posts. The objective of the classification is to understand the opinion of a population on a particular subject. The machine-learning approach uses a sentiment classifier to differentiate a negative sentiment from a positive one. The sentiment classifier algorithms need training data to build a predictive model to classify new data. Since training data are labelled manually, the process of building a predictive model is known as supervised learning. Naive Bayes, Support Vector Machines, Maximum Entropy, Random Forest, and Logistic Regression are some of the most used sentiment classifiers (Kotze and Senekal 2018, 3).

2.2 Sentiment Analysis of Turkish Tweets

Since English is widely spoken, many sentiment analyses of English texts exist in the literature. However, the analysis of Turkish texts is still a research area open to development (Tokçaer 2021, 1). There is some research regarding the sentiment analysis of Twitter posts in Turkish. Aimed at measuring the accuracy of machine learning classifiers, some research has been conducted. For instance, Ayan et al. analyzed and labelled 162,000 tweets as either positive or negative. They employed Ridge Regression and Naive Bayes classifiers to train the data set. The results demonstrate that Ridge Regression yields a higher f1 score, a measurement of a test’s accuracy in binary classification, which is 96.9%. while Naive Bayes yields 95.4 %. Additionally, the results highlight that the Ridge Regression requires less training time (Ayan et al. 2019, 500).

Kumaş analyzes 1600 positive and 1600 negative Turkish tweets related to diverse subjects. The analysis demonstrates that among Naive Bayes, K-Nearest Neighbours, Logistic Regression, Decision Tree, and Support Vector Machines classifiers, the latter proves to be the best algorithm based on the f1 score. Kumaş does not employ any n-gram techniques to investigate their impact on the performance of classifiers. (Kumaş 2021, 1)

Onan contributes to the literature by scrutinizing the effects of n-gram feature representation schemes on the success of machine learning classifiers. An n-gram means a sequence of n words. For instance, “machine learning” is a bigram and “machine learning approach” is a trigram sequence. In the context of sentiment analysis, using frequent n-grams as features instead of single words may result in a better classification. Onan applies Naive Bayes, Support Vector Machines, and Logistic Regression classifiers to 5,300 positive and 5,300 negative tweets. The classifiers' performances are measured with 1-gram, 2-gram, and 3-gram feature representation schemes. Although the correct classification rates obtained with Naive Bayes and Logistic Regression classifiers are close to each other, the highest correct classification rate is obtained by the Naive Bayes algorithm with 1-gram and 2-gram schema. (Onan 2017, 9)

Coban et al. set forth an extended form of the research of Onan and evaluates the rate of success of Naive Bayes, Multinom Naive Bayes, Support Vector Machines, and K-Nearest Neighbours for the bag of words and n-gram feature representation schemes. Research suggests that the Multinom Naive Bayes algorithm is the most successful combined with the N-gram scheme. (Coban et al. 2015,4)

In the context of analyzing the public opinion of mainstream society about minorities by employing the machine learning approach, Turkish literature offers limited research. Most of the research is qualitative, therefore, the analysis of big data remains inefficient.

Chapter 3: Methodology

3.1 Data Collection

Since there is no former research related to the Roma minority concerning sentiment analysis in Turkey, the dataset was created using Twitter API. The keyword list for tweet extraction includes *çingen, çingen, çingene, çingeneler, çingene, çingeneler, çingane, çingane*. The word ‘çingene’ means ‘gypsy’ in the Turkish language and has a pejorative meaning. *Çingen* and *çingane* are the different forms of *çingene* while *çingeneler* is its plural. At the end of the collection and pre-processing process, 723 tweets have been assigned sentiment manually. Tweets have been collected using Python 3.10.4 and Tweepy 4.8.0 packages between April 24, 2022, and May 17, 2022. Pandas 1.4.2 was used to convert tweets into a data frame.

3.2 Data Pre-processing

Tweets include many characteristic aspects, and they are unstructured data. They contain specific characters like "@", "#", and irrelevant links for sentiment analysis. Therefore, they should be processed before conducting the analysis. In this context:

- User tags "@", hashtags "#", and links "URLs" were removed from tweets using regular expressions.

  Retweets "RT" and duplicate tweets were removed from the data set.

  Punctuation characters, numbers, and emojis has been removed from tweets.

  The content has been converted into lower case letters.

  Terms that include less than three characters has been removed from tweets.

  Tweets that were not posted in the Turkish language has been detected by the Python langdetect 1.0.9 package and they were removed from the data set.

  Stop words are the words in a language that do not convey a meaning themselves. In this research, NLTK 1.7.0 package for natural language processing with Python has been used to remove Turkish stop words.

3.3 Tokenization and Stemming

Tokenization means breaking raw text into smaller chunks to facilitate the interpretation of the meaning of the text by analysing the sequence of words. In the context of analysing tweets, tokenization means breaking each tweet into words and storing them in an iterable, mostly a list structure. To tokenize Turkish tweets, the word tokenize module of the NLTK package has been used in the research.

Stemming is the reduction of words into their base form to reduce computation. To make it clearer, since there is no significant difference between a word and its plural form in terms of deriving the sentiment, reducing them into their roots improves the efficiency of sentiment analysis. For example, there is no difference among the words ‘management’, ‘manager’, or ‘to manage’ in terms of conveying the sentiment. In this research, the Turkish Stemmer 1.3.0 package has been utilized for stemming.

3.4 Text Vectorization

Text vectorization, also known as text representation, is the process of transforming a text into numerical form so that the data can be used to train a machine learning model like logistic regression. There are numerous techniques to convert text into a numerical form. In this research, positive/negative word frequencies, bag-of-words (BoW), and term frequency-inverse document frequency (TF-IDF) techniques has been applied to represent text.

The positive/negative word frequencies technique is a simplistic method for representing the text. The first step is to create a corpus that consists of unique words present in tweets collected. Then each word is assigned a positive frequency and a negative frequency feature. Positive frequency is the number of appearances of a unique word in tweets conveying positive sentiment while the negative frequency represents the count of that word in negative tweets. After features, in other words, frequencies, are assigned to each word, the sentiment of a new tweet can be estimated comparing the positive and negative word frequencies of that tweet. If the sum of positive word frequencies is greater than the sum of negative word frequencies in that tweet, the positive label can be assigned.

Like the positive/negative word frequencies technique, features of bag-of-words model are extracted based on unique words of a tweet corpus. A unique value is assigned to each unique word, and that unique value constitutes the feature. After features are extracted, each tweet is converted into a binary vector storing the information of whether a feature is present in a tweet or not. Vectorized tweets can be transferred to a features and vectors matrix. The dimension of the matrix is the number of unique words times the number of tweets. For instance, if the bag includes 1,000,000 unique words and 10,000 tweets are in the corpus, the dimension of the feature matrix is 10,000,000. The matrix obtained is the training data for a specific machine learning model. The Count Vectorizer function of sci-kit-learn 1.0.2 is used to apply BoW.

TF-IDF is a sophisticated text vectorization technique compared to bag-of-words and positive/negative word frequencies. Because it puts emphasis on the words that are more significant in terms of conveying a sentiment, this makes TF-IDF one of the widely used presentation techniques. Like BoW, features consist of unique words present in the corpus and for each feature, TF and IDF scores computed. TF, term frequency, represents the frequency of a unique word in a tweet while IDF, inverse document frequency, is the logarithm of the total number of tweets divided by the total number of tweets that a unique word is present. The more a word is present in a tweet corpus, the less the IDF score is for this word. Therefore, it is less significant for sentiment derivation.

3.5 The Classifier and Model Training

Logistic Regression is a classification algorithm used in machine learning to predict a binary/categorical outcome based on some input data, independent variable/variables. It calculates the probability whether an event occurs or not. Therefore, the dependent variable, is always between 0 and 1.

The logistic function can be used to model a binary dependent variable and it is expressed as $1 / 1 - e^{(-t)}$. The t in the equation represents the linear combination of explanatory variables. In the context of sentiment analysis of tweets, t computed by the elements of the vectors obtained after the text representation process. Based on t value of each tweet, probability of a tweet's sentiment being positive is calculated using logistic function. If the probability is greater than 0.5 the tweet is considered as positive.

To apply logistic regression, the Logistic Regression function of sci-kit-learn 1.0.2 package is used. 80% of tweets is used as training dataset, when 20% is used as testing dataset. Three different models are built using different text vectorization methods mentioned earlier.

3.6 The Model Performance Measures

The accuracy of the model changes based on the text vectorization methods applied. The accuracy-score function of sci-kit-learn package is employed to calculate it for each method. It simply calculates the proportion of sum of true negative and true positive, to the sum of true negative, true positive, false negative, and false positive.

True negative means that the model estimated a sentiment as 0 when its real value is 0. Similarly true positive means that the model estimated a sentiment as 1 when its real value is 1. On the other hand, false negative means that the model estimated the sentiment's value as 0 while the true value is 1, and false positive occurs when the model's estimate is 1 and the true value is 0. To evaluate the performance of each model in estimating the true sentiments, see the confusion matrices below.

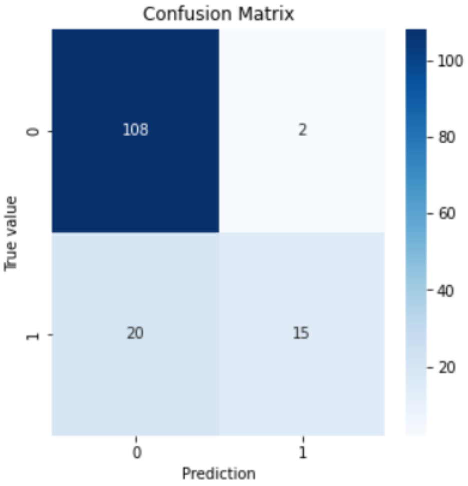

Figure 1: Confusion matrix for positive/negative word frequencies. (Graph by author)

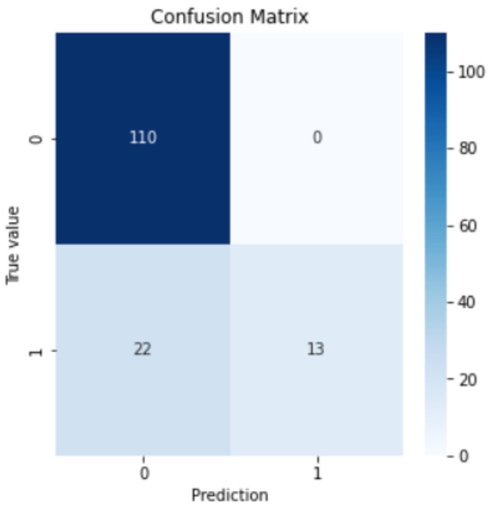

Figure 2: Confusion matrix for BoW.
(Graph by author)

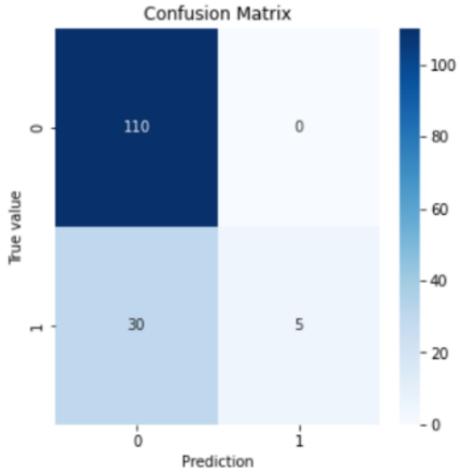

Figure 3: Confusion matrix for TF-IDF.
(Graph by author)

From the confusion matrices, it can be clearly seen that models are quite efficient estimating negative sentiments. However, they perform badly in estimating positive sentiments, and this results in Type II error. The reason behind is that the dataset is not balanced. The proportion of positive tweets is not enough to train the model properly in estimating positive sentiment.

For the accuracy scores, see Table 1 below.

TABLE 1. The Model Accuracy Score for Each Text Vectorization Technique

Text Vectorization Method	**Model Accuracy Score**
Positive/Negative Word Frequencies	84.83%
Bag-of-Words	84.83%
Term Frequency – Inverse Document Frequency	79.31%

Source: (Table by author)

Although TF-IDF produces better results in general, it is observed that bag-of-words and positive-negative word frequencies have proved more efficient in deriving sentiment from the tweets.

Chapter 4: The Generalization from Data

In this section, the aim is to test statistically whether the Turkish tweets related to gypsies convey a negative sentiment in general. Based on this, the null hypothesis asserts that the Turkish tweets related to “gypsy” convey a positive sentiment while the alternative hypothesis claims the opposite.

The tweets analysed and positive tweets were represented as 1 while negative ones were represented as 0. Therefore, the average of sentiment values corresponds to the proportion of positive tweets in the dataset. Also, this proportion gives the sample mean. Based on this, a hypothesis testing can be performed to make an inference about the population mean of sentiment values. In this case, the population represents all tweets posted related to “gypsy” in Turkey in 2022.

The null hypothesis asserts that the mean is greater or equal to 0.5 which indicates that the population has a positive sentiment in general. On the other hand, alternative hypothesis asserts a mean smaller than 0.5 which signifies the negative sentiment of the population.

To perform the hypothesis testing the confidence level is determined as 95 per cent. As the sentiment values can be either 0 or 1, the highest possible variance 0.25 of sentiment values is used to determine the confidence interval. The sample size is 723 while the corresponding z value is approximately 1.96.

TABLE 2. The Parameters and CI for Hypothesis Testing

Parameter	**Value**
n (sample size)	723
σ^2 (highest possible variance)	0.25
confidence level	95%
z (confidence level value)	1.96
CI (confidence interval)	[0.463, 0.536]
p value	< 0.00001

Source: (Table by author)

The 95 per cent confidence interval is calculated as [0.463,0.536]. This means that the mean of the population lies between 0.463 and 0.536 with 95 per cent probability in case we accept the null hypothesis. Under the assumption of the null hypothesis a sample with 0.23 mean is of very low probability. That means the probability of selecting a sample whose mean is smaller or equal to 0.23 is of a chance smaller than 0.0001. Therefore, the null hypothesis must be rejected and there is overwhelming statistical evidence demonstrating that most Turkish tweets discussing “gypsies” express negative feelings.

Chapter 5: Summary of Results

Approximately 77 per cent of the tweets has been found to convey negative sentiments while the 23 per cent convey positive sentiments.

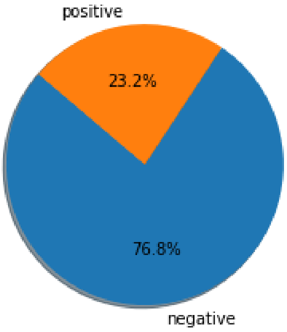

Figure 4: The percentages of positive and negative tweets.
(Graph by author)

The tweets contain a plethora of pejorative phrases related to the Roma population in Turkey. They solidify the stereotypical perception of the Roma and contribute to the creation of hate speech.

5.1: The Analysis of Tweets with Positive Sentiment

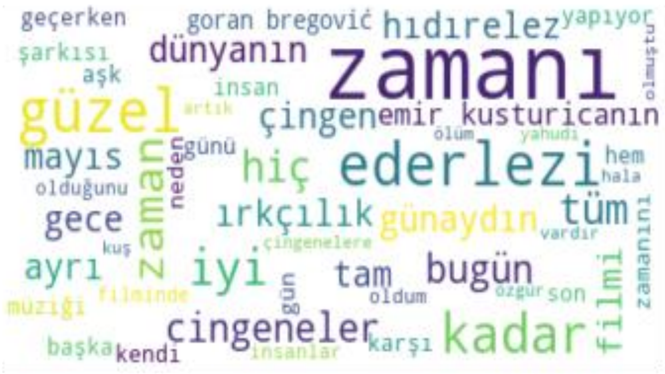

Figure 5: The frequently used words in positive tweets.

(Image by author)

The words in the figure include *aşk* ('love'), *mayıs* ('may'), *ırkçılık* ('racism'), *ederlezi* ('Roma celebration of spring'), *gece* ('night'), *iyi* ('good'), *insan* ('human'), *kuş*('bird'), *zaman* ('time'), *şarkı* ('song'), *özgür* ('free'), *müzik* ('music').

- Tweets conveying positive sentiment constitute 23.2 percent of the sample. Figure 2 displays the most common words in positive tweets. The mainstream identifies the Roma with art, specifically music, and the film *Time of the Gypsies* is frequently mentioned in positive tweets. Additionally, criticism of racism is a recurring theme in the positive tweets.

5.2 The Analysis of Tweets with Negative Sentiment

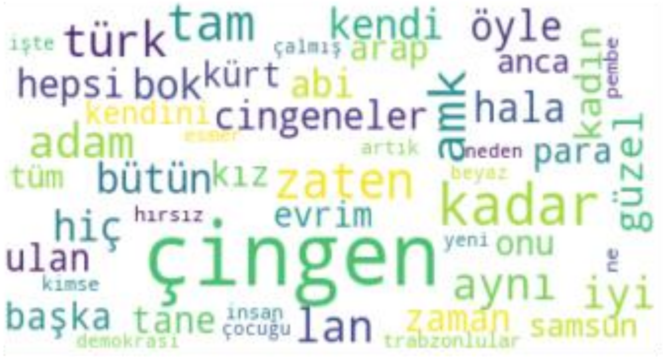

Figure 6: The frequently used words in negative tweets.

(Image by author)

The words in the figure include hepsi ('all'), aynı ('same'), anca ('just'), hiç ('nothing'), kadın ('woman'), kız ('girl'), Türk ('Turkish'), Kürt ('Kurdish'), Arap ('people from the Middle East''), demokrasi ('democracy'), para ('money'), çalmak ('to steal'), adam ('man'), pembe ('pink').

The tweets conveying negative sentiment constitute 76.8 percent of the sample. Figure 3 displays the most frequently used words in negative tweets.

- The words *kürt* ('Kurdish') and *arap* ('people from the Middle East') demonstrate that the mainstream puts ethnic minorities in the same category, and they convey a negative sentiment. The Kurds are the largest minority group in Turkey with the greatest political power while the *arap* encompasses refugees in Turkey.
- The words *hırsız* ('thief'), *para* ('money') and *çalmak* ('to steal') show that the Roma are seen as potential criminals, and they are a threat to society.
- Numerous slang words are common in tweets like *ulan* ('bub'), *lan* ('bub'), *bok* ('shit'), *amk* ('pussy'). These words indicate the disgust and rage held towards the Roma and exemplify the hate speech on social media platforms.
- The common usage of the word *kadın* ('woman') and *kız* ('girl') in negative tweets indicates the objectification and eroticization of the Roma woman by the mainstream. Females are usually affiliated with black magic and prostitution.
- Lastly, the existence of words *beyaz* ('white') and *esmer* ('brunette') and may show that the Roma are discriminated against based on their appearance.

Chapter 6: Limitations

- Since tweets are collected between April,4 of 2022 and May,17 of 2022, the dataset includes a limited number of tweets to be analysed. There were 723 tweets in total. Additionally, the dataset does not involve a proportional number of positive and negative tweets. The number of negative tweets was 555, while it was 168 for the positive tweets. This results in Type II statistical error in estimates of models. However, in a more balanced dataset where the number of positive and negative tweets are close to each other, Type II error can be eliminated.
- The research does not consider different Gypsy groups in Turkey. Although there are three main Gypsy groups in Turkey, which are Rom, Dom, and Lom, they are considered as one group in this study. However, they have adopted the culture of different ethnic groups based on their geographic locations, therefore, their lifestyle and identity perception may differ from each other.

Chapter 7: Conclusions

The tweets demonstrate that mainstream Turkish society has a mostly negative perception of the Roma. The word "çingene" is used in the tweets to belittle and insult an individual or a community frequently. The Roma are seen as uneducated, impoverished, and quarrelsome. They are mostly affiliated with crimes like robbery and kidnapping.

Considering that this study is conducted based on Turkish tweets posted from Turkey, it can be a basis for future studies that aim to investigate the situation of the Roma or other marginalized groups in the country through sentiment analysis on social media. The results obtained by the analysis may constitute a basis for policymaking, media representation, and civil society initiatives. Particularly, understanding the perception of the mainstream is of high importance to break down the prejudices and produce knowledge via media channels.

References

Aytuğ, Onan. "Twitter mesajlari üzerinde makine öğrenmesi yöntemlerine dayali duygu analizi." Yönetim Bilişim Sistemleri Dergisi 3, no. 2 (2017): 1-14.

Çoban, Önder, Barış Özyer, and Gülşah Tümüklü Özyer. "Sentiment analysis for Turkish Twitter feeds." In 2015 23rd Signal Processing and Communications Applications Conference (SIU), pp. 2388-2391. IEEE, 2015.

Hakan, Alp. "Çingenelere yönelik nefret söyleminin Ekşi Sözlük'te yeniden üretilmesi." *Ankara Üniversitesi İLEF Dergisi* 3, no. 2 (2016): 143-172.

Kolukirik, Suat. "The perception of Gypsies in Turkish society." *Roma Rights Quarterly* 3 (2007): 31-36.

Kotzé, Eduan, and Burgert Senekal. "Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa." TD: The Journal for Transdisciplinary Research in Southern Africa 14, no. 1 (2018): 1-11.

Kumaş, Enes. "Türkçe Twitter Verilerinden Duygu Analizi Yapılırken Sınıflandırıcıların Karşılaştırılması." Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi 2, no. 2 (2021): 1-5.

Nguyen, Thu T., Hsien-Weng Meng, Sanjeev Sandeep, Matt McCullough, Weijun Yu, Yan Lau, Dina Huang, and Quynh C. Nguyen. "Twitter-derived measures of sentiment towards minorities (2015–2016) and associations with low birth weight and preterm birth in the United States." Computers in human behavior 89 (2018): 308-315.

Tokcaer, Sinem. "Türkçe Metinlerde Duygu Analizi." Yaşar Üniversitesi E-Dergisi 16, no. 63 (2021): 1516-15