Sentiment Analysis for the Social Media: A Case Study for Turkish General Elections

Conference Paper · April 2017

DOI: 10.1145/3077286.3077569

4 authors:

Elif Uysal
Konya Food and Agriculture University

Semih Yumusak
KTO Karatay University

Kasim Oztoprak
Konya Food and Agriculture University

Erdogan Dogdu
Angelo State University

Sentiment Analysis for the Social Media: A Case Study for Turkish General Elections

Elif Uysal
KTO Karatay University
Konya, Turkey
elif.uysal@ogrenci.karatay.edu.tr

Semih Yumusak
KTO Karatay University
Konya, Turkey
semih.yumusak@karatay.edu.tr

Kasim Oztoprak
KTO Karatay University
Konya, Turkey
kasim.oztoprak@karatay.edu.tr

Erdogan Dogdu
Georgia State University (adjunct)
Cankaya University
Ankara, Turkey
edogdu@cankaya.edu.tr

ABSTRACT

The ideas expressed in social media are not always compliant with natural language rules, and the mood and emotion indicators are mostly highlighted by emoticons and emotion specific keywords. There are language independent emotion keywords (e.g. love, hate, good, bad), besides every language has its own particular emotion specific keywords. These keywords can be used for polarity analysis for a particular sentence. In this study, we first created a Turkish dictionary containing emotion specific keywords. Then, we used this dictionary to detect the polarity of tweets that are collected by querying political keywords right before the Turkish general election in 2015. The tweets were collected based on their relatedness with three main categories: the political leaders, ideologies, and political parties. The polarity of these tweets are analyzed in comparison with the election results.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Natural Language Processing

Keywords

Social media, sentiment analysis, political tweets

1. INTRODUCTION

Social media has proven to be a game changer in politics [6, 7], it contains information that can be used for predicting the future streams [5]. In social media, people express their opinions by using short and informal phrases. Although the opinions are not expressed in complete and uniform sentence structures, it is possible to detect the mood and the emotion of the words and phrases [8]. Based on this idea, we searched for people's political tendencies in Twitter just before the 2015 general elections. We categorized the politics-related tweets in Turkish language as `positive, negative, and neutral`, which were first collected and filtered by using parties' names, leaders, abbreviations and ideologies[1]. For this study, we used SentiStrength tool and Zemberek [2] library for natural language processing (NLP) operations in Turkish language. Complete experimental setup, results and the customized Turkish dictionary is available in the project repository[2].
We review the related work in Section 2. Then, the methods we used for sentiment analysis and tweet collection are described in Section 3. Experimental results are presented in Section 4, and we conclude in Section 5.

2. RELATED WORK

Sentiment analysis for the social media is studied in a number of works for different languages. Similar to our study, Vural [9] proposes a sentiment-focused Web crawling and sentiment analysis in Turkish. For sentiment analysis in Turkish, Zemberek's morphological analyzer and SentiStrength tool are used after customizing the lexicon files for Turkish language. They used the same tools we use in this study to evaluate the sentiment analysis framework for online movie reviews, hotel reviews, and political news in Turkish language. In our study, we use political Turkish tweets for sentiment analysis by creating a new Turkish sentiment dictionary. Akbas [1] proposes a method for aspect-based opinion mining for Turkish tweets. Zemberek library is used for capturing negation, stemming and spell checking. For the sentiment strength detection in Twitter, a combination of lexical and machine learning approaches are used. A sentiment-based text representation of tweets is used and SentiStrength is configured for Turkish language. [4] on the other hand proposes an emotion analysis for Turkish texts. In their study, ISEAR dataset and Turkish fairy tales dataset are used for analysis. Complement Naive Bayes, Naive Bayes and SVM are used for classification and Zem-

ACM SE '17, April 13-15, 2017, Kennesaw, GA, USA

DOI: http://dx.doi.org/10.1145/3077286.3077569

[1] https://tr.wikipedia.org/wiki/T%C3%BCrkiye%27deki_siyasi_partiler_listesi
[2] https://github.com/uysalelif/TurkishSentimentAnalysis

berek library is used for morphological analysis.
A similar study is proposed by [3] for English using political tweets for Irish general elections in February 2011. Similarly, they classify political tweets as positive, negative, and neutral for Irish general elections. In that study, tweets were collected using Twitter search API according to parties and parties' leaders by performing query-based search. For the sentiment polarity of a word, subjectivity lexicon and SentiWordNet 3.0 datasets are used. For the 3-class classification of tweets, a naive lexicon-based approach and supervised machine learning are used.

3. METHODS

Sentiment analysis methods can be used to classify people's opinions about subjects on the social media or the Web as positive, negative, and neutral statements. In other words, sentiment analysis in this case is about understanding the polarity on a subject. We used SentiStrength library for the sentiment analysis of tweets, and used Zemberek framework, an open source Turkish natural language processing software, for word suggestions, suffix removal, and the detection of negative verbs with negation suffixes.
Algorithm 1 explains our method for the sentiment analysis of tweets. Tweets are read and parsed first. A number of NLP processing operations are done on tweets. Misspelled words are corrected, word stemming is done, and negative words are identified. Finally, a sentiment score is calculated using our Turkish sentiment dictionary.

Algorithm 1: Sentiment Analysis of Tweets

```
while ∃ next tweet do
    for each word in tweet do
        if misspelled(word) then
            replace word with the best suggestion word;
        end
        remove suffixes from verb and noun stem, and detect
          negative verb that includes '-me,-ma' (negative
          suffix);
    end
    Calculate the sentiment score of tweet and write the
      result to output file;
end
```

3.1 Libraries

Zemberek Library: Zemberek[3] is an open source, platform independent natural language processing (NLP) framework designed for Turkish language. Zemberek library provides some well-known NLP operations such as spell checking, morphological analysis, stemming, word suggestion, word construction, extracting syllables [2]. In this work, Zemberek library is used for word suggestions, suffix removal from nouns, verb stemming, and negative verb detection [9] [1] [4].

SentiStrength Library: SentiStrength[4] is a sentiment analysis library, which supports the English language as default; however, can be customized to other languages by changing the input files. This library scores short texts with two types of strengths, namely negative strength and positive strength. Negative strength ranges from -1 (not negative) to -5 (extremely negative) and positive strength ranges from 1 (not positive) to 5 (extremely positive).
In order to use this library for the Turkish language, we customized the library for the Turkish language, which is similarly implemented by [9] and [1]. Different from these studies, we created our own dictionary by using the Turkish dictionary and combined the input files with the ones provided by [9]. We created a dictionary containing 1543 emotion words and 120 emoticons, and used in the SentiStrength library as input.

3.2 Tweet Collection

We used a Python-based library[5] to query tweets and export these to a comma separated (csv) file. Tweets are collected by querying political party names, abbreviations, ideologies and leaders of four major political parties that are represented in the general election. The collection process was limited to a 1-week time interval, which is the week before the election until the election day. We collected the tweets by using a query search within those boundaries. After the tweet collection step, the collected political tweets are used for analyzing the polarization of tweets that is explained in section 4.

4. EXPERIMENTS

In this section, the experimental setup is explained and the results for the analysis of the collected tweets are discussed.

4.1 Experimental Setup

We have used the Zemberek library for the sentiment analysis in the Turkish language. Using the Zemberek library, first we applied a word suggestion process, then we removed the suffixes from noun/verb stems and detected the negative verbs that include negation suffixes. For example, in Table 1, the word '*sirketin*' is suggested as '*şirketin*'. Then, '*şirket*' (company) root word is obtained by removing the suffix. For handling negative verbs in Turkish, '*negativeword*' is written as a replacement of verbs that have negation suffixes in tweets and it is analyzed with score -5 in the process of sentiment analysis, which is illustrated in Table 2. Table 2 shows negative verbs that include negation suffixes and when detecting these words, they were replaced by the '*negativeword*' for sentiment analysis. However, handling negative verbs did not work very well because of Turkish language structure. For example, in Table 3 the word '*konuşma*' refers to '*speech*' in English and this word is a

Table 1: example tweet and its analysis

original tweet	Eski seçimlerdeki performanslari ile başarılı buldugum 5 sirketin son anket ortalamalari: AKP %42.0 CHP %26.7 MHP %16.1 HDP %11.1 Koalisyon
analyzed tweet	**3 -1** eski seçim performans ile **başarı[3]** bul 5 şirket son anket ortalama akp %42 .[sentence: 3,-1] 0 chp %26 .[sentence: 1,-1] 7 mhp %16 .[sentence: 1,-1] 1 hdp %11 .[sentence: 1,-1] 1 koalisyon [sentence: 1,-1] [result: max + and - of any sentence][overall result = 1 as pos>-neg]
English translation	Polling averages of the companies that I found successful in the previous elections: AKP %42.0 CHP %26.7 MHP %16.1 HDP %11.1 Coalition

[3]https://code.google.com/archive/p/zemberek/
[4]http://sentistrength.wlv.ac.uk/
[5]https://github.com/Jefferson-Henrique/GetOldTweets-python

Table 2: A sample tweet for '*negativeword*'

original tweet	Ülkücülük bir ufuktur başı dik olmayanlar göremezler
analyzed tweet	**1 -5** ülkü bir ufuk baş dik **negativeword[-5] negativeword[-5]** [-1 consecutive negative words] [sentence: 1,-5] [result: max + and - of any sentence][overall result = -1 as pos<-neg]
English translation	Nationalism is a horizon, but those who are not upright cannot see.

noun. The Zemberek library perceives this word as a verb with a negation suffix and this tweet was marked as negative. In fact, this tweet does not include any sentiment, thus should not be marked as negative. In this study, those words could not be excluded from the results, which may slightly decrease the accuracy of the tagging process.

Using the SentiStrength library, we classified tweets with

Table 3: An example of an incorrect evaluation of a tweet

original tweet	3) Devlet bahçeli **konuşma** yapıyor++
analyzed tweet	**1 -5** 3)devlet bahçe **negativeword[-5]** yap [sentence: 1,-5] [result: max + and - of any sentence][overall result = -1 as pos<-neg]
English translation	Devlet bahçeli making a speech

Table 4: List of sample emotion words in Turkish

edepli (*decent*)	3	gurur (*pride*)	2
mükemmel (*excellent*)	4	görkemli (*splendid*)	4
mutsuzluk (*unhappiness*)	-2	harika (*marvelous*)	3
faydasız (*useless*)	-2	kaygı (*anxiety*)	-4
gergin (*nervous*)	-3	keder (*sorrow*)	-4

input files sentimentally. In order to use this library for the Turkish language, we edited the dictionary to include Turkish emotion keywords. Table 4 is an example of some emotion words in Turkish with scores. If the result of an analyzed tweet is *pos<-neg*, this tweet is taken *negative*, if result of an analyzing tweet is *pos>-neg*, this tweet is labeled as *positive* and in other situations they are labeled as neutral.

Other than the emotion words, SentiStrength also has scores for emoticons. In Table 5, an example of an analyzed tweet with emoticons is listed.

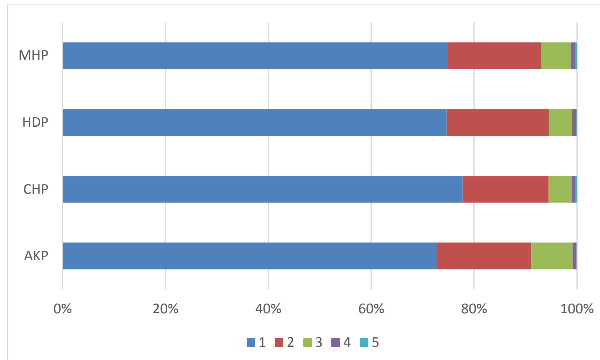
4.2 Results and Analysis

We retrieved 100669 tweets with the above mentioned criteria. We grouped the collected tweets to three categories, namely political party leaders, abbreviations and ideologies (e.g. using query keywords like "bozkurt", "milliyetçi hareket partisi", "adalet ve kalkınma partisi", "radikal demokrasi", "chp", etc.). We have filtered out 7016 of the tweets that include some of the search keywords in for example the user names (e.g. '*bozkurt*' is a political keyword, but there are user names that includes this word).

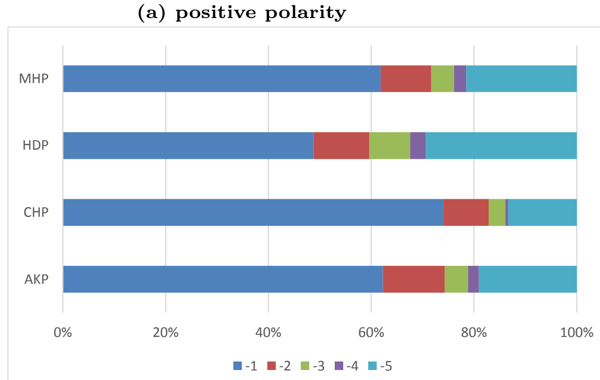
In the analysis of tweets, positive polarity score ranges from 1 to 5, negative polarity score ranges from -1 to -5 and the overall polarity score ranges from -4 to 4. Figure 1 represents the polarity of all parties with the scale of positive and negative in terms of the parties' names. *Score of 1* has the highest portion and *score of 5* has the lowest portion in the range in Figure 1a. *Score of -1* has the highest portion and *score of -4* has the lowest portion in the range in Figure 1b.

Table 5: An example of a tweet with emoticons

original tweet	@GokayGuler gerçi ben mhp ye atıyorum o ayrı :)
analyzed tweet	**2 -1** @gokayguler gerçi ben mhp ye at o ayrı **:) [1 emoticon]** [sentence: 2,-1] [result: max + and - of any sentence][overall result = 1 as pos>-neg]
English translation	@GokayGuler Anyway, I voted for mhp by the way :)



(a) positive polarity

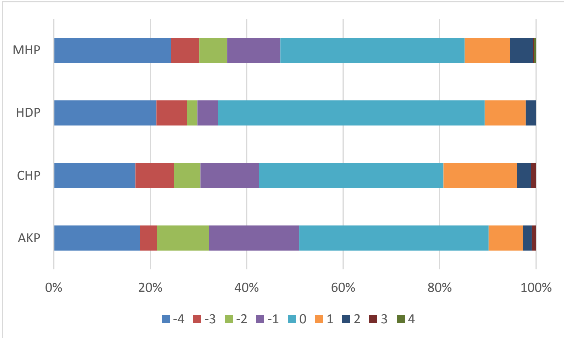


(b) negative polarity

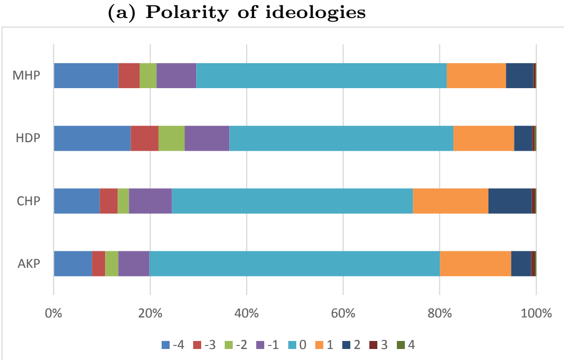
Figure 1: Negative and positive polarity of tweets related to party names

Figure 2 displays the sentiment distribution of tweets with the mention of political party ideologies, the names of party leaders and the names of political parties. In Figure 2a, '*mhp*' has the highest portion and '*hdp*' has the lowest portion with the score of 4. In addition to that, '*chp*' has the lowest portion with the score of -4 and '*mhp*' has the highest portion with the score of -4. As a result, the ideological keywords does not provide any significant correlation with the election results.

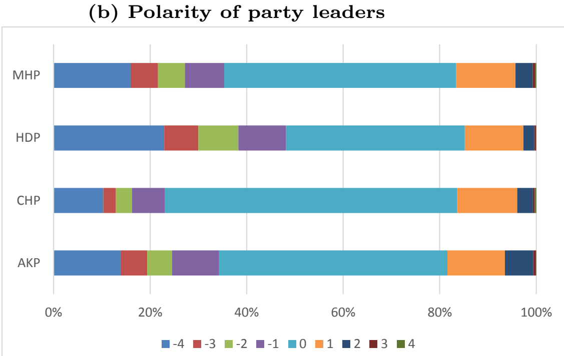
In Figure 2b, '*mhp*' has the lowest portion with the score of 4, but for the other parties, there is no difference for the score of 4. Besides that, '*mhp*' has the highest portion and '*akp*' has the lowest portion with the score of -4. As a result, the polarity of the tweets with the leader names positively correlates with the election results for both negative and positive ranges.



(a) Polarity of ideologies



(b) Polarity of party leaders



(c) Polarity of party names

Figure 2: analysis of tweets based on three categories

In Figure 2c, '*chp*' has the highest portion and '*hdp*' has the lowest portion for the score of 4. Also, '*chp*' has the lowest portion and '*hdp*' has the highest portion with the score of -4. As a result, the tweets including party names are determinative clearly in both sides.
Consequently, '*akp*' is the first party in terms of the leading polarity. On the other hand, '*chp*' is the first party with respect to the party names. In fact, '*akp*' was the winner of the elections held in June 2015. According to this result, the polarization of leaders has the closest results by analyzing the Turkish political tweets. Furthermore, the second ranked party was '*chp*', '*mhp*' was the third, followed by '*hdp*' as the last.[6]
Finally, Table 6 lists the polarity results of the tweets for all of these three categories (ideologies, leaders, and party names).

[6] https://en.wikipedia.org/wiki/Turkish_general_election,_June_2015

Table 6: categorization of tweets based on party's ideologies, leaders and names

	ideologies	leaders	names
positive	257	5482	10029
negative	684	7862	27442
neutral	597	14934	26366
total	1538	28278	63837

5. CONCLUSION

In this study, we analyzed people's sentiment tendency using tweets mentioning political opinions during Turkish elections in 2015. We classified 93,653 Turkish political tweets that include the names of political parties, the names of the leaders of parties, and the ideological keywords of political parties. For this categorization, we used a variety scale of sentiment analysis. We obtained better results in the categorization of leaders and party names.
As for future work, we will study sentiment analysis on search engine queries. The number of sentiment words in the dictionary will be extended and the software framework will be refined to handle negative verbs to increase the accuracy of the polarity scores.

6. REFERENCES

[1] E. Akbaş. Aspect Based Opinion Summarization With Turkish Tweets. Master's thesis, Bilkent University, 2012.
[2] A. A. Akın and M. D. Akın. Zemberek, An Open Source NLP Framework for Turkic Languages. *Structure*, 2007.
[3] A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. In *Association for Computational Linguistics*, 2013.
[4] Z. Boynukalin. Emotion analysis of turkish texts by using machine learning methods. *MS Thesis, Middle East Technical University*, 2012.
[5] C. Chung and K. Austria. Social Media Gratification and Attitude toward Social Media Marketing Messages: A Study of the Effect of Social Media Marketing Messages on Online Shopping Value. *Proc. of the Northeast Business & Econ. Assoc.*, pages 581–586, 2010.
[6] P. N. Howard and M. R. Parks. Social Media and Political Change: Capacity, Constraint, and Consequence. *Journal of Communication*, 62(2):359–362, 2012.
[7] M. Lim. Clicks, Cabs, and Coffee Houses: Social Media and Oppositional Movements in Egypt, 2004-2011. *Journal of Communication*, 62(2):231–248, 2012.
[8] R. McKerlich, C. Ives, and R. McGreal. Measuring use and creation of open educational resources in higher education. *International Review of Research in Open and Distance Learning*, 14(4):90–103, 2013.
[9] A. G. Vural, B. B. Cambazoglu, and P. Karagoz. Sentiment-focused web crawling. *ACM Transactions on the Web (TWEB)*, 8(4):22, 2014.