Structured Learning Based Turkish Sentiment Analysis
Yapılandırılmış Öğrenmeye Dayalı Türkçe Duygu Analizi

Oguz Ulgen and Arif Selcuk Ogrenci
Graduate School of Science and Engineering, Kadir Has University, Istanbul, Turkey
email : {oguz.ulgen, ogrenci}@khas.edu.tr

***Abstract*—Sentiment analysis is highly popular topic to identify people's opinions through the social media, forums and other websites. There are an abundance of opinions on internet and analysing those opinions would have many benefits for both private and public sectors. Research has evolved looking on tweets for mining opinions and for the classification of the tweets as positive, negative or neutral in its sentiment. In this research, Turkish tweets are used for sentiment extraction where a two layer neural network is used as the pattern recognition system. The supervised training of this system is based on structured learning. As a conclusion, structured learning seems to be helpful in pattern recognition to classify tweets and mining the opinions. However, it is evident that further research in data processing and training methodology is necessary to obtain reliable sentiment analysis results.**

***Keywords*—sentiment analysis, twitter, neural network, pattern recognition, structured learning**

***Özetçe* —Duygu analizi, sosyal medyada, forumlarda ve diğer internet sitelerinde insanların fikirlerini belirlemek için sıkça kullanılmaktadır. İnternette çok sayıda fikir bulunmakta ve bu fikirlerin analiz edilmesi özel sektör ve kamu sektörü için birçok faydayı beraberinde getirmektedir. Araştırmalar, fikirleri toplamak için tweetleri kullanma ve bu tweetleri pozitif, negatif ve nötr olarak sınıflandırma yönünde evrildi. Bu araştırmada, örüntü tanıma sistemi olarak çift katmanlı sinir ağı kullanılırken duyguları almak için Türkçe tweetler kullanıldı. Bu sistemin denetlenen eğitimi yapılandırılmış öğrenime dayanmaktadır. Sonuç olarak yapılandırılmış öğrenme, duyguları sınıflandırmak ve görüşleri incelemek için model tanımada yardımcı olur gibi görünüyor. Bununla birlikte, güvenilir bilgi analizi sonuçlarını elde etmek için veri işleme ve eğitim metodolojisinde ileri araştırmaların yapılması gerektiği açıktır.**

***Anahtar Kelimeler*—duygu analizi, twitter, sinir ağları, örüntü tanıma, yapısal öğrenme**

I. Introduction

Aristotle defines human as a "political animal" which implies that people are meant to live in a community. Rapid development on technologies has made the whole globe as one small community. One of the results of this technology and globalization is social media. With the help of social media, now, people express their opinions, ideas, concerns and interests in an easy manner. This creates an abundance of information about the community and the people. This opportunity creates space for sentiment analysis, especially on Twitter [1]. By checking the context and the content of a tweet, its polarity can be decided. A tweet can be either positive, negative or neutral [2-3].

Research about sentiment analysis is carried out in many fields such as getting the costumer's opinions about a product [4-5], predicting election outcomes[6], reviews about movies and books [7] and also for some other types of data analysis [8]. Although there are many papers about social media sentiment analysis most of them are for the English language and there are not many found for the Turkish language. There are a lot differences between English and Turkish, so different methods have to be employed analyze sentiment in different languages.

Sentiment analysis basically has four main steps: data collection, feature extraction, feature selection and classification. For data collection, internet sites and their RSS feeds or some APIs can be used. Feature extraction generally is done by methods like stemming, tokenization, lemmatization etc. After the feature extraction, features are selected for analysis and for that purpose methods such as NLP, clustering and statistical analysis can be used. After all these steps classification can be done by some supervised or unsupervised learning approach. The rest of this paper is organized as follows: Section 2 will give information about related Works; Section 3 describes the methodology; in Section 4 results will be given and Section 5 will conclude the paper.

II. Related Work

A. Sentiment Analysis

In the literature there are many papers about the research in sentiment analysis. One of the early examples of sentiment analysis is the work of [2]. In this paper, we can see that sentiment analysis is defined and future works are mentioned. Later sentiment analysis become very famous and attractive to the researchers. In [9] the reasons of the topics' popularity are considered. It is mentioned that, sentiment analysis has a wide application field almost in all of the domains of interest. Owing to the fact that it was a new topic, it was offering brand new problems that are never studied before. Moreover, a huge amount of opinion is provided by social media or other websites (e.g. forums, blogs, reviews), so it has become so much easier to gather data than before.

A human can only read, consider and gather a limited amount of sentiments about a topic in the web. On the other hand, automated systems can gather data and process them in a pace of thousands of time faster than a human. Therefore, in real-life applications sentiment analysis systems are considered as highly attractive. Main applications of sentiment analysis are based on marketing, public relations and political campaigns [9]. A lot of big companies like Microsoft (Azure), Google (Google Cloud Natural Language API), and Hewlett-Packard

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

(Haven) have been working on sentiment analysis and using the results in their favor.

Apart from the industrial uses, many researchers are interested in the topic of sentiment analysis for academic purposes. In [10], research is carried out on sentiment analysis in multiple languages on web forums. In [11], Zhang and Skiena considered trading strategies from blogs and news sentiments. In [12], characterization of social relations is done via NLP sentiment analysis.

Neural networks can be deployed for sentiment categorization by using unsupervised (Self Organizing Maps) or supervised (pattern recognition) techniques. Data mining, web mining and information gathering are open to applications of sentiment analysis.

B. Twitter Sentiment Analysis

In 2006 Twitter has been opened to use by the public. Until 2008 it was not very popular. After a while, people noticed that they can comment whatever is on their mind with the restriction of 140 characters and mention whoever they want. This could be a politician, singer, actress etc. After they saw that they can reach these people, an increasing number of people started to write what they think. This created an abundance of opinions and it was very attractive for sociologists and also for many big companies to analyze those tweets. After all, they do not have to go out and do some survey anymore, answers were already there. With the help of computer softwares it got easy to collect thousands of tweets in just a few seconds. However, interpretation of those tweets required still a great deal of work. Automated sentiment analysis systems would allow easy data processing for tweets in comparison to the blogs or forums, because there are a maximum of 140 characters used in a single tweet.Some examples of sentiment analysis based on tweets are given in [13-15].

C. Turkish Twitter Sentiment Analysis

Like in many other languages, some research has also been done for sentiment analysis on twitter for the Turkish language. An early example has been given about Turkish politics in [16]. In this work, different machine learning methods are employed and their performance has been compared. Another research was about classification of Turkish tweets with the help of LDA (Latent Dirichlet Allocation) [17]. Most of the research work has utilized the NLP tool called Zemberek [18] for language processing. It is a free, open source Natural Language Processing framework that can be used for spell checking, lemmatization, stemming, word suggestion etc. With the help of Zemberek language processing and feature selection is provided.

III. Methodology

In this paper, NLP and sentiment analysis techniques will be used for Turkish language. As an NLP tool Zemberek NLP will be used. With the help of it, it is possible to stem, tokenize and analyze most words. After preprocessing Neural Network Toolbox of MATLAB will be used to implement pattern recognition system model.

basari	basla	biz	Fenerbahce	mac	yasa

Table I. Created Array for The Example

Tweet #	abd	arka	basari	basla	basta	bir	...
# 1	0	1	0	0	0	0	...
# 2	1	0	1	0	0	0	...
# 3	0	0	1	1	0	0	...

Table II. Matrix represents which words are appeared in which tweets

A. Data Preprocessing

When the data are collected, tweets are full of misspelled words and some meaningless data for the research. The following parts of tweets are removed:

- Punctuation and symbols
- Mentioned usernames on tweets
- Words whose length is less than 3 characters
- Hashtags

After the removal of the abovementioned parts with the help of Zemberek each word in the tweet is analyzed and checked if there is any misspelling. For the misspelled words the first suggested word is chosen. The set of orrectly spelled words and the suggested corrections are stemmed and alphabetically sorted. These stems are passed to an array and this process is repeated for each tweet in the data.

For example, the original tweet is ”Yasasin #fenervsmadrid baslyor. Basarilar Fenerbahcem, bu mac bizim ! @fbbasketbol” which is translated in English as ”Hurray! #fenervsmadrid is strting. Good Luck Fenerbahce, this game is ours ! @fbbasketbol”. Firstly, ”#fenervsmadrid”, ”.,!” and ”@fbbasketbol” are removed from this tweet. After that, misspelled words are processed: the suggested word ”basliyor” is selected instead of ”baslyor” and it is chosen. Each word is stemmed as follows: ”yasa, basla, basari, Fenerbahce, mac, biz” and it is passed to an array like in Table I. The whole process can be seen in the flowchart in Figure 1.

B. Creating Vocabulary

For each tweet the data preprocessing is done. After each word is stemmed and passed to an array, the corresponding tweet is mapped for the array and it is passed to a matrix whose dimensions are (number of tweets) X (number of words). In this matrix, if a word appears in a tweet, a 1 is placed in the corresponding position of the matrix. Similarly, a 0 is placed in the matrix if the word does not appear. The final outcome of this procedure is depicted in Table II.

C. Sentiment Assignment

For each tweet, the ultimate sentiment decider is a human, so it has a subjective evalation of some person. One can also

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

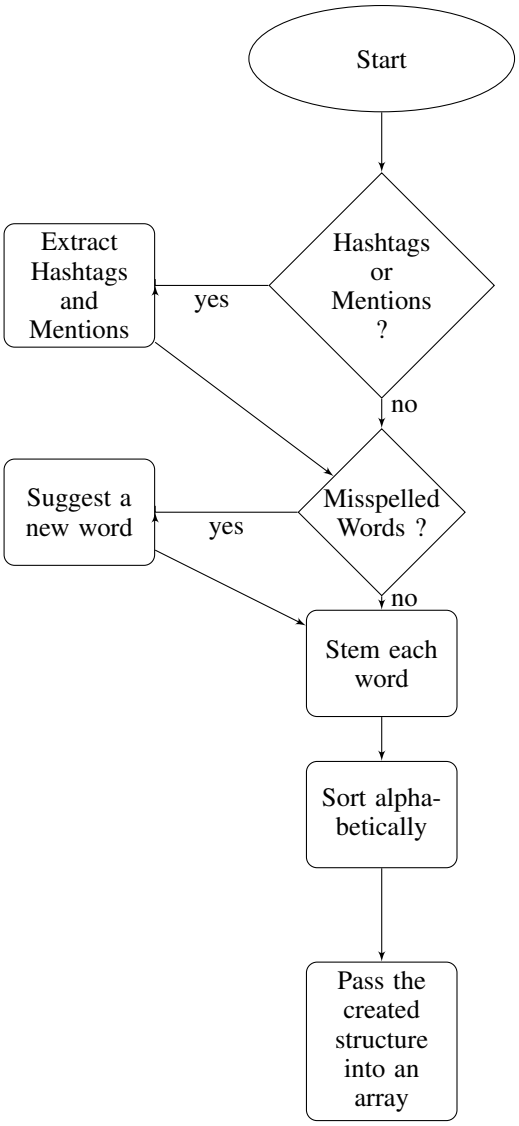

Figure 1. Flowchart of Data Preprocessing

argue that the concept of sentiment has to be subjective by its nature. The following output categories are assigned for each tweet. If a tweet contains negative polarity "0 0 1", if a tweet contains positive polarity "1 0 0" else if a tweet contains neutral polarity "0 1 0" is passed to a 2D Matrix whose dimensions are (number of tweets) X 3. This scheme characterizes an encoding by 3 outputs. In this research we have also investigated the case with 2 outputs only: if a tweet is negative "0 1", if a tweet is non-negative "1 0" is passed to a 2D Matrix (number of tweets) X 2. Both of these encodings are used to compare system performance.

D. Training The Network

A neural network is trained with the tweet and sentiment matrix. Feedforward pattern recognition network is used in the MATLAB Neural Network Toolbox. For pattern recognition, two-layered neural network is used. In the hidden layer, the system has sigmoid transfer function and in output layer softmax has been utilized. The network is trained with scaled

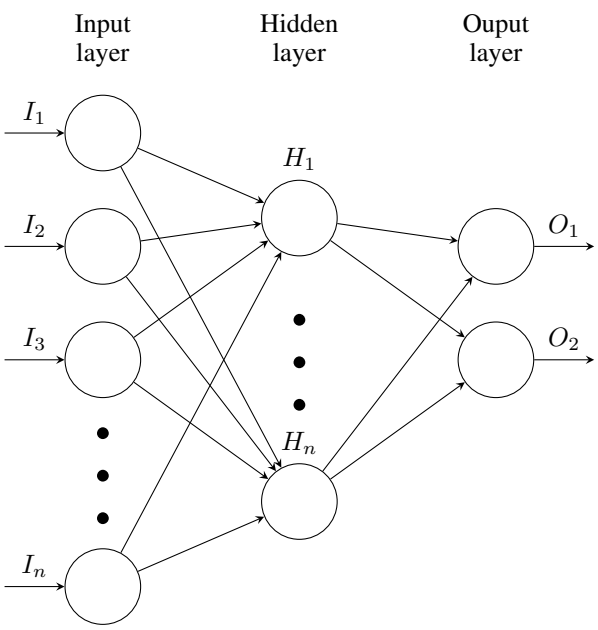

Figure 2. Neural Network Model with 2 outputs

conjugate gradient backpropagation algorithm. The model of the neural network used is shown in Figure 2.

IV. RESULTS

For different number of tweets different number of features (or words) are obtained shown in Table III. It is seen that as the number of tweets increase the increase of number of features slows down. Since the number of words is limited, after some amount number of tweets will have no or little effect on number of features.

Number Of Tweets	Number Of Features
100	424
250	763
500	1245
700	1537

Table III. NUMBER OF FEATURES CHANGE RESPECT TO NUMBER OF TWEETS

To train the network 700 tweets chosen either negative, positive or neutral. Out of 700, 540 of the tweets are selected for training, 70 for validation and 70 tweets for testing the training. System has 3 outputs that shows the polarity of a tweet. The resulting Test Confusion Matrix with positive, negative and neutral results are shown in Figure 3. It is seen that system classified positive tweets with 52% success while negative tweets are categorized correctly by 81.5%. However, success rate for neutral tweets is only 5.6%. As a result, the general success rate is 51.4%.

After completing the training with 3 outputs, the second scenario is implemented. In the second scenario, 2 output system is implemented. Again the total sample size of 700 is divided into training, validation and test sets made up of 540, 70, and 70 tweets respectively. The resulting Test Confusion Matrix with negative and non-negative results is shown in

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

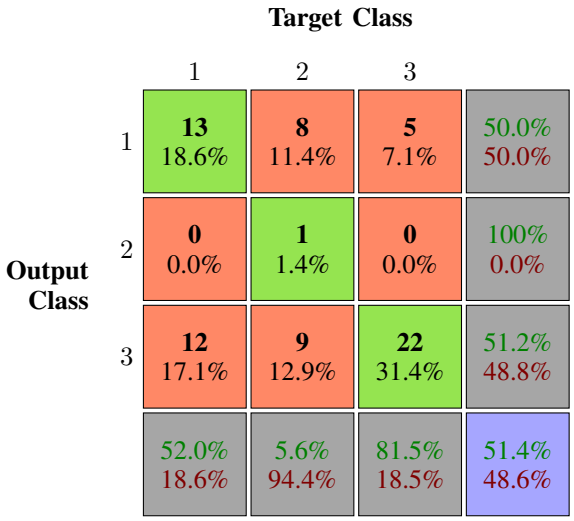

Figure 3. Test Confusion Matrix for 700 Tweets with 80-10-10 Rate with 3 Outputs

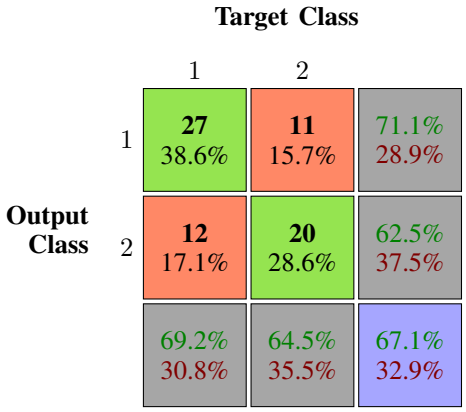

Figure 4. Test Confusion Matrix for 700 Tweets with 80-10-10 Rate with 2 Outputs

Figure 4. As seen in the figure 69.2% and 64.5% are the success rates for non-negative and negative tweets respectively. General system classification rate is 67.1%. In [13], where similar methods are used, when Twitter API used for the data gathering, the result was only around 50%.

Results show that when the number of output encoding is 2, the success rate to classify tweets is better than using 3 outputs for encoding. This might help if only negative content is to be found in a tweet.

V. Conclusion

As a conclusion, for structured learning 2-output system results better than 3-output system because it is hard for system to distinguish neutral sentiment rather than positive and negative. If the system tries to classify only negative and non-negative than it is seen that almost 70% of success rate is obtained. Since this was a preliminary work, future works will be studied. For future work, better feature selection methods will be implemented. Different Neural Networks will be tried and the results will be compared.

References

[1] A. Pak and P. Paroubek, ""Twitter as a corpus for sentiment analysis and opinion mining."" in LREC, vol. 10, 2010, pp. 1320–1326.

[2] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, ""Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques"" in Third IEEE International Conference on Data Mining, 2003 (ICDM 2003), Melbourne, Florida, USA, 2003

[3] W. Rong, B. Peng, Y. Ouyang, C. Li, and Z. Xiong, "Semi-supervised Dual Recurrent Neural Network for Sentiment Analysis" in Dependable, Autonomic and Secure Computing (DASC), 2013 IEEE 11th International Conference on, Dec 2013, pp. 438-445.

[4] W. Wang, "Sentiment Analysis of Online Product Reviews With Semisupervised Topic Sentiment Mixture Model," in Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, vol. 5, Aug 2010, pp. 2385-2389.

[5] R. L. de S. Santos, R. F. de Sousa, R, A. L. Rabelo, R. S. Moura , "An Experimental Study Based on Fuzzy Systems and Artificial Neural Networks to Estimate The Importance of Reviews About Product and Services" in International Joint Conference on Neural Networks (IJCNN), 2016, Vancouver, BC, Canada

[6] M. Anjaria and R. Guddeti, "Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning" in Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, Jan 2014, pp. 1-8.

[7] M. Neethu and R. Rajasree, "Sentiment Analysis in Twitter Using Machine Learning Techniques" in Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference on, July 2013, pp. 1-5.

[8] Zamahsyari and A. Nurwidyantoro, "Sentiment Analysis of Economic News in Bahasa Indonesia Using Majority Vote Classifier" in International Conference on Data and Software Engineering (ICoDSE), 2016, Denpasar, Indonesia, Indonesia, 2016.

[9] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[10] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums", ACM Transactions on Information Systems, 26(3), pp. 1–32, 2008.

[11] W. Zhang, S. Skiena, "Trading Strategies to Exploit Blog and News Sentiment", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, 2010.

[12] G. Groh, J. Hauffa, "Characterizing Social Relations Via NLP-based Sentiment Analysis", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011.

[13] B. Duncan and Y. Zhang, "Neural networks for sentiment analysis on twitter," in Cognitive Informatics Cognitive Computing (ICCI*CC), 2015 IEEE 14th International Conference on, July 2015, pp. 275–278.

[14] Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[15] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment Analysis on Tweets for Social Events," in Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, June 2013, pp. 557-562.

[16] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, pp. 174–180, 2012.

[17] O. Coban and G. T. Ozyer, "Sentiment classification for Turkish Twitter feeds using LDA" in 24th Signal Processing and Communications Applications Conference (SIU), Zonguldak, Turkey, 2016

[18] A. A. Akin and M. D. Akin, "Zemberek, An Open Source NLP Framework For Turkic Languages," Structure, vol. 10, pp. 1–5, 2007.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.