Developing Turkish Sentiment Lexicon for Sentiment Analysis Using Online News Media

Fatih Sağlam[1], Hayri Sever[2] and Burkay Genç[3]
[1]Business Administration, [2]Computer Engineering, [3]Institute of Population Studies
[1]Turkish Military Academy, [2,3]Hacettepe University
Ankara, Turkey
[1]fsaglam@kho.edu.tr, [2]sever@hacettepe.edu.tr, [3]burkay.genc@hacettepe.edu.tr

***Abstract*—Internet is a very rich resource of documents that need to be analysed to extract their sentimental values. Sentiment Analysis which is a subfield of Natural Language Processing discipline focuses on this issue. The existence of sentiment lexicons in their own language is a very important resource for scientists studying in sentiment analysis field. Since many studies of sentiment analysis have been conducted on text written in English language, developed methods and resources for English may not produce the desired results in other languages. In Turkish, a rich sentiment lexicon does not exists, such as SentiWordNet for English. In this study, we aimed to develop Turkish sentiment lexicon, and we enhanced an existing lexicon which has 27K Turkish words to 37K words. For quantifying the performance of this enhanced lexicon, we tested both lexicons on domain independent news texts. The accuracy of determining the polarity of news written in Turkish has been increased from 60.6% to 72.2%.**

Keywords—sentiment analysis; Turkish; GDELT; Turkish Sentiment Lexicon; SentiWordNet

I. INTRODUCTION

Internet users, with Web 2.0, are also provided with the opportunity to participate in content development beyond surfing. This causes a large amount of unstructured data that is growing rapidly. Retrieving information and using it in decision making process from this data have become a focus of research for many scientists. In this context, one of the research fields is Sentiment Analysis that tries to reveal the attitude and emotional state of a given text [1].

Since there is an immediate need of processing the opinionated web contents originating from social media, sentiment analysis or opinion mining became a rapidly emerging research field [2].

As Pang et al. [3] stated, various terminologies can be used to describe a specific subject. Same challenge also exists in sentiment analysis. "Opinion Mining", "Subjectivity Analysis", "Review Mining", "Appraisal Extraction" and "Affective Computing" phrases have been used similarly in the literature by some scientists [4]. In this paper, we prefer to use "Sentiment Analysis".

One of the main goals of sentiment analysis is to exercise "Sentiment Polarity Classification" which is used to obtain the semantic polarity (positive, negative or neutral) of a text. Lexicon based and machine learning based approaches are employed in order to extract the semantic polarity automatically from a given text [5], [6], [7], [8], [9].

Since most of the studies on sentiment analysis are based on text written in English language, current sentiment lexicons and corpora are mostly in this language [10], [11]. Moreover, scientists usually evaluate the performance of their studies on English texts. In contrast, there are very few studies done in other languages, such as Turkish. Existing sentiment analysis methods developed for English rarely have productive outcome when it comes to Turkish due to the fact that Turkish is an agglutinative language. As Kaya et al. [10] stated, statistical methods have poor performance when it comes to morphologically rich languages such as Turkish.

Some significant differences between English and Turkish languages have been given by Vural et al. [12]. These differences can be summarized as follows:

- Turkish is an agglutinative language, i.e., root words can be extended by many suffixes to produce new meanings. Some examples are given in Table I.

TABLE I. EXAMPLES OF TURKISH WORDS

Word	Suffixes	English Meaning
oku		read
okudum	oku-du-m	I read
okuyorum	oku-yor-um	I am reading
okuyabilirim	oku-yabilir-im	I can read
okuyamayabilirdim	oku-yamayabilir-di-m	I might not have been able to read

978-1-5090-4320-0/16/$31.00 ©2016 IEEE

- The added suffixes may change the polarity of a root word. An example is given in Table II.

TABLE II. AN EXAMPLE OF CHANGING POLARITY OF A ROOT WORD

Word	Suffixes	English Meaning	Semantic Polarity
tehlikeli	tehlike-li	dangerous	Negative polarity
tehlikesiz	tehlike-siz	safe	Positive polarity

- In Turkish, negation should be handled very carefully compared to the negation in English, because words can be negated by suffixes hidden within the word. An example is given in Table III.

TABLE III. AN EXAMPLE OF A NEGATION WORD IN TURKISH

Word	Suffixes	English Meaning
saldırdı	saldır-dı	attacked
saldırmadı	saldır-**ma**-dı	did not attacked

Just as English has polarity lexicons like SentiWordNet [13], other languages including Turkish require similar resources. Due to all these reasons, the main motivation of this study has been the development of a Turkish sentiment lexicon.

The rest of the paper is organised as follows: In section 2, we provide a brief survey of previous work on Turkish sentiment analysis and polarity detection. Section 3 gives our proposed approach for a creating Turkish sentiment lexicon. This lexicon has been evaluated on a manually labeled Turkish corpora in section 4. In section 5, we discuss the results of our study and propose future work.

II. RELATED WORKS

Ucan [14] in his thesis produced Turkish Sentiment Dictionary by translating from English to Turkish. This lexicon has 27K Turkish words with assigned polarity scores. He used SVM (Support Vector Machine) method with this lexicon on movie review corpora to determine its performance. This lexicon, called hereafter as SWNetTR, constitutes the basis of our study.

Erogul [15] studied sentiment analysis in Turkish in his thesis. He gathered movie reviews from popular related web sites (http://rec.arts.movies.reviews, http://rottentomatoes.com and http://beyazperde.com) and performed analysis with SVM. He also analysed the effects of part-of-speech (POS) information of words and negation suffix on the sentiment of the reviews. In his study he did not develop any comprehensive Turkish sentiment lexicon.

KAYA et al. [10] studied sentiment analysis in Turkish political news. They constructed a dataset which consists of political news from articles in different Turkish news sites. The dataset used in this study was domain dependent (political) and also conducted with machine learning based approach.

The remarkable features of the previous studies related to sentiment analysis can be summarised as follows:

- Machine learning based approach is often used.
- Instead of using a comprehensive sentiment lexicon, an opinion-words dataset which is composed of most frequently used words in related domain is used.
- Experiments are usually run on domain dependent texts.
- The number of sentiment analysis studies conducted in Turkish language is quite low.

The only study which we encounter regarding sentiment analysis for Turkish news belongs to KAYA et al. [10], however it is also domain dependent (political news).

III. PROPOSED APPROACH

In this paper, we have used lexicon based approach where the collective polarity of a document or a sentence is the sum of polarities of the individual words or phrases.

Main steps of our proposed approach is given in Fig. 1. As seen in this figure, we start with a large database of Turkish news pages on the web whose URLs are taken from the GDELT database. We parse these HTML pages to obtain raw news text. We, then, use the Zemberek framework to obtain the roots of the words seen in these texts. We assign a score to each word using the polarity values from the GDELT database. This results in the SWNetTR-GDELT lexicon of around 14000 unique Turkish words. We then extend the existing SWNetTR lexicon with around 10000 unique words that exists in SWNetTR-GDELT but not in SWNetTR. We call this new lexicon the SWNetTR-PLUS. We finally test the new lexicon and report results.

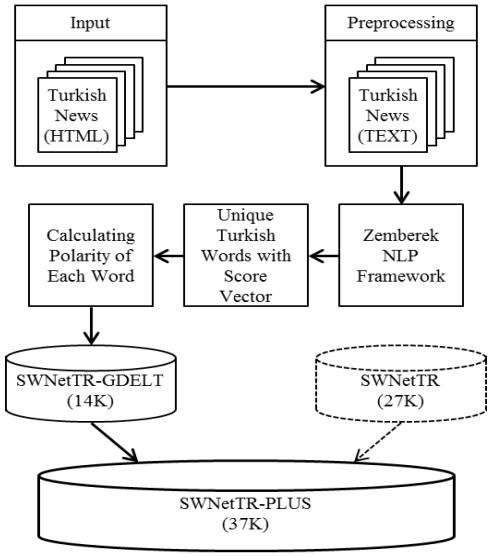

Fig. 1. Main steps of proposed approach.

Table IV outlines the lexicons mentioned in this paper, providing the corresponding abbreviations and descriptions.

TABLE IV. LEXICONS ABBREVIATIONS AND DESCRIPTIONS

Abbreviation	Description
SWNetTR	A Turkish SentiWordNet containing 27K words developed by Ucan et al. [14].
SWNetTR-GDELT	A Turkish SentiWordNet, constructed by us, containing 14K words compiled from Turkish news at GDELT.
SWNetTR-PLUS	A Turkish SentiWordNet, combining SWNetTR and SWNet-GDELT containing 37K words.
MLTC	Manual Labeled Turkish Corpora developed by us.

A. SWNetTR

Ucan [14] in his thesis constructed a polarity scored Turkish sentiment lexicon which contains 27K Turkish words. To do that, he translated English words in SentiWordNet to Turkish via different online translation sites (http://www.tureng.com, http://www.zargan.com, http://tr.bab.la and Google Translate API). To obtain reliable results, he also used some novel methods such as serial translating and parallel translating. However, he did not translate the English words which have neutral polarity to Turkish. Since the polarity of a word may vary between languages, translating the all words in English to Turkish may be a more effective approach, regardless of semantic polarity. It seems to be a shortcoming of his study.

In this study, our aim is to enhance SWNetTR and enrich it with new words.

B. SWNetTR-GDELT

SWNetTR-GDELT is a Turkish sentiment lexicon constructed within the scope of this study. We make use of datasets provided by the GDELT Project (http://gdeltproject.org).

GDELT (Global Database of Events, Language and Tone) is defined by its website as a very large and comprehensive open database of human society. It monitors and analyses the world's news media in many formats over 100 languages, translating them to English where necessary. They extract information such as events, actors, locations, themes and semantic tones from news websites on the Internet. The outcome of this analysis is released in 15 minute periods on Google BigQuery which is a cloud-based analytics database. GDELT Project conforms to Big Data 5V (Variety, Volume, Velocity, Veracity, Value) concepts. Although GDELT is launched in 2013, its content dates back to 1979.

GDELT presents essentially two main datasets: "Events" and "GKG (Global Knowledge Graph)". GKG contains the actual graph connecting all persons, organisations, locations, emotions, themes, counts, events, and sources together each day into a single network structure and captures the cultural narratives that envelope the global information stream. In our study, we used GKG dataset to obtain URLs of news and their tone values.

Briefly, GDELT converts unstructured text to structured data. In a short period, it has become a popular data source for academic community, and it is our study's main source of data.

In order to create SWNetTR-GDELT, first we randomly selected a total of 100000 Turkish news URLs, half of which has positive and the other half has negative tone scores taken from GDELT GKG dataset. GDELT's tone value is equivalent to the document-level score in sentiment analysis domain. We fetched and parsed the news text at these URLs. At the time of the study, 82912 of these web pages were online. Next, we preprocessed the fetched text to exclude irrelevant terms and symbols such as HTML tags. Each remaining word was then stemmed by Zemberek framework which is an open-source NLP framework for Turkic languages[1]. Zemberek tries to find the possible roots of a given word via morphological parsing approach. Some examples of the Zemberek framework are given in Table V.

TABLE V. ZEMBEREK FRAMEWORK EXAMPLES

Input (Word)	Output (Root Form of Word)
oynamak (to play)	oyna
oynadı (he played)	oyna
oyuncu (player)	oyun
oyuncak (toy)	oyuncak

The tone value provided by GDELT for a news article is then assigned to each root word produced by Zemberek as a polarity score. At the end of this preprocessing, we obtained 14K unique Turkish words along with their score vector.

An example of the score vector of the word "*keder*" (sorrow) is given at Table VI.

TABLE VI. SCORE VECTOR OF "*KEDER*" WORD

Word	Frequency (*f*)	Polarity Score (*d*)
keder	1	-3.1145
keder	4	-5.2008
keder	1	-4.4423
keder	2	-2,9854

[1] http://code.google.com/p/zemberek

To calculate the polarity score of each word, we computed the weighted mean (1) of the corresponding tone vector.

$$S_w = \frac{\sum_{i=1}^{n} (d_i * f_i)}{\sum_{i=1}^{n} f_i} \quad (1)$$

n: Total number of documents that contain the word w.

d_i: Polarity score of d (document-level)

f_i: The number of occurrences of the word w within the document.

S_w: Computed the polarity score of word w.

As a result, we have developed a Turkish lexicon-level sentiment lexicon called SWNetTR-GDELT which has semantic score values between -1.0 and 1.0, binary polarities -1 or +1, POS information. SWNetTR-GDELT contains 14023 unique words. In essence, this method is the bag-of-words approach, in which a text is represented as an unordered collection of words.

C. SWNetTR-PLUS

When we compared SWNetTR-GDELT with SWNetTR, we realised that there are 10253 Turkish words in SWNetTR-GDELT that does not exist in SWNetTR. We computed ranking rate for these extra words as follows (2):

$$R_{rate} = \frac{\sum_{i=1}^{n} r_i}{n(n-1)/2} \quad (2)$$

n: Total number of words in SWNet-GDELT.

r_i: Rank of extra word which is not in SWNetTR.

R_{rate}: Ranking rate.

Calculation result is 0.735, which means that many words used in Turkish news are not in SWNetTR.

SWNetTR-PLUS (Fig. 2) is formed by combining SWNetTR and SWNetTR-GDELT.

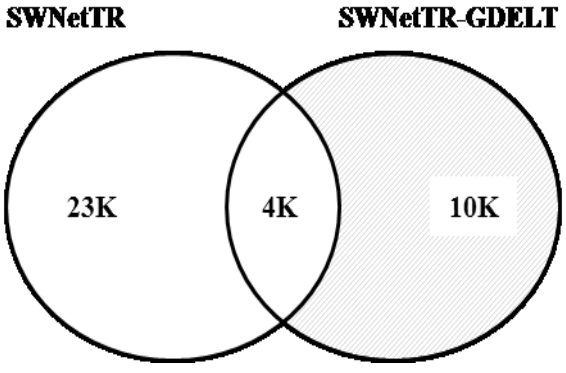

Fig. 2. Number of words in SWNetTR-PLUS, and the corresponding originating lexicons.

D. MLTC : Manual Labeled Turkish Corpora

To evaluate the performance of SWNetTR-PLUS, we need an annotated corpus in Turkish. Due to the lack of such a resource, we had to develop an annotated corpus by ourselves.

The MLTC was constructed to evaluate the performance of SWNetTR-PLUS. A sample of 500 news texts was chosen, such that the average polarity was nearly neutral.

The collected data was annotated by three native speakers of Turkish, and calculated *mode* value for each text is assigned as the polarity (positive or negative). As a result, MLTC test dataset is produced.

IV. Experimental Evaluation

Chi-Squared statistical test has been done for the evaluation of performance phase. Test results prove that it is statistically meaningful.

A. Confusion Matrix

Confusion matrix is a specific table layout that contains information about actual and predicted classifications [16]. It also allows visualisation of the performance of an algorithm or an approach which is used for classification. Confusion matrix layout is shown in Table VII. In this table, TN stands for True Negative, FN stands for False Negative, FP stands for False Positive and TP stands for True Positive.

TABLE VII. Confusion Matrix Layout

		PREDICTED CLASS	
		Neg.	Pos.
ACTUAL CLASS	Neg.	*TN*	*FP*
	Pos.	*FN*	*TP*

To interpret the results of confusion matrix, there are some terms and calculation methods, such as accuracy, recall and precision. The “Accuracy” term refers to the proportion of the total number of predictions that were correct. To calculate the performance of SWNetTR and SWNetTR-GDELT lexicons on MLTC corpora, we used the accuracy formula given at (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

B. MLTC vs SWNetTR

The confusion matrix which is prepared for the comparison of MLTC and SWNetTR is given in the Table VIII. Test result reveals that the accuracy of the SWNetTR is 60.6%.

TABLE VIII. SWNETTR – MLTC CONFUSION MATRIX

		SWNetTR		
		-1,00	1,00	Total
MLTC	-1,00	66	169	235
	1,00	28	237	265
	Total	94	406	500

C. MLTC vs SWNetTR-PLUS

The confusion matrix which is prepared for the comparison of MLTC and SWNetTR-PLUS is given in the Table IX. Test results reveal that the accuracy of the SWNetTR-PLUS is 72.2%.

TABLE IX. SWNETTR-PLUS – MLTC CONFUSION MATRIX

		SWNetTR-PLUS		
		-1,00	1,00	Total
MLTC	-1,00	187	48	235
	1,00	91	174	265
	Total	280	220	500

According to the results given in Table VIII and Table IX, SWNetTR-PLUS lexicon has higher accuracy over SWNetTR lexicon. This outcome is expected since the word count of SWNetTR-PLUS is higher than that of SWNetTR. Another remarkable result is a significant difference in value between True Negative (TN) values in aforementioned tables. It means that our SWNetTR-PLUS lexicon has more words with negative polarities than the ones in SWNetTR. It is also an expected result due to the fact that the media has more negative news, and these news texts were used as a source of the construction of SWNetTR-GDELT.

V. CONCLUSION AND FUTURE WORK

We provide two novelties in this study. First, we have constructed polarity scored Turkish sentiment lexicon using online news media. Second, we test existing lexicons on domain independent texts, which were previously only tested on domain dependent texts. Therefore our work is the first one to perform an evaluation of Turkish news which is completely domain independent.

The 27K words SWNetTR polarity lexicon is upgraded to 37K words. In order to evaluate the performance of this upgraded lexicon, MLTC has been constructed. Both the performance of 27K SWNetTR and 37K SWNetTR-PLUS has been compared by the use of MLTC.

The results show that accuracy of the polarity classification performance done by SWNetTR, which is composed by translations from English, is 60.6% .

SWNetTR-PLUS polarity lexicon which is developed within the scope of this study performed with an accuracy of 72.2% .

We aim to improve the capacity and performance of SWNetTR-PLUS as a future work. For this purpose we plan to work on the concept of negation in Turkish language, POS and position information. We also plan to work further on common words that exist in both SWNetTR and SWNetTR-GDELT to test whose polarities for these words provide a better lexicon.

REFERENCES

[1] G. Vinodhini and R.M. Chandrasekaran, “Sentiment analysis and opinion mining: a survey,” International Journal 2.6, 2012.

[2] A. Nasser, K. Dinçer, H. Sever, “Investigation of feature selection problem in sentiment analysis for Arabic language,” Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkey, April 3-9, 2016.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no 1-2, 2008, pp. 1-135.

[4] R. Piccard, “Affective computing,” MIT Press, 1997.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, “Lexicon-based methods for sentiment analysis,” Association for Computational Linguistics, 37(2), 2011, pp. 267-307.

[6] F. Akba, A. Ucan, E.A. Sezer & H. Sever “Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews,” In 8th European Conference on Data Mining, 2014.

[7] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” In Proceedings of LREC, vol. 6, 2006, pp. 417-422.

[8] C. Kaushik and A. Mishra, “A scalable, lexicon based technique for sentiment analysis,” International Journal in Foundations of Computer Science & Technology (IJFCST), vol .4, no.5, 2014, pp. 35-43.

[9] C. Mate, "Product aspect ranking using sentiment analysis: A survey," International Research Journal of Engineering and Technology, vol.03, issue 01, 2015, pp. 126-127.

[10] M. Kaya, G. Fidan and I.H.Toroslu, "Sentiment analysis of Turkish political news," Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol.01, IEEE Computer Society, 2012.

[11] E. Cambria, B. Schuller, Y. Xia, C. Havasi, “New avenues in opinion mining and sentiment analysis,” IEEE Intelligent Systems, vol. 2, 2013, pp. 15-21.

[12] A.G. Vural, B.B. Cambazoglu, P. Senkul and Z.O. Tokgoz, “A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish,” In Computer and Information Sciences III, Springer London, 2013, pp. 437-445.

[13] B.N. Raju, C. Naikodi and L. Suresh, “Sentiment analysis of product attribute using social media,” International Journal of Engineering Research, vol.5, 2016, pp. 808-813.

[14] A. Ucan, “Automatic sentiment dictionary translation and using in sentiment analysis,” (Master’s thesis, Hacettepe University, Ankara, Turkey, 2014), retrieved from www.cs.hacettepe.edu.tr/lisansustu/files/aucanmaster.pdf

[15] U. Erogul, “Sentiment analysis in Turkish,” (Master’s thesis, Middle East Technical University, Ankara, Turkey, 2009), retrieved from https://etd.lib.metu.edu.tr/upload/12610616/index.pdf

[16] R. Kohavi and F. Provost, “Glossary of terms. Machine Learning,” 30(2-3), 1998, pp. 271-274.