The Impact of NLP on Turkish Sentiment Analysis

E. Yıldırım
İTÜ Computer Eng. &
Informatic Faculty

F. S. Çetin
İTÜ Computer Eng. &
Informatic Faculty

G. Eryiğit
İTÜ Computer Eng. &
Informatic Faculty

T. Temel
İTÜ Computer Eng. &
Informatic Faculty

Abstract

Sentiment analysis on English texts is a highly popular and well studied topic. On the other hand, the research in this field for morphologically rich languages is still in its infancy. Turkish is an agglutinative language with a very rich morphological structure. For the first time in the literature, this paper investigates and reports the impact of the natural language preprocessing layers on the sentiment analysis of Turkish social media texts. The experiments show that the sentiment analysis performance may be improved by nearly 5 percentage points yielding a success ratio of 78.83% on the used data set.

1 Introduction

Sentiment analysis has become a very popular re-search area because of needs to track and man-age population tendency. Many companies to-day works on this area in order to meet cus-tomer expectations and demands.

Gönderme ve kabul tarihi: 17.09.2014-25.10.2014

Social micro-blogging platforms (e.g. Twitter and Facebook) offer an opportunity to get huge amount of eas-ily accessible and processable data. Users of micro-blogging platforms write about their per-sonal lives, their own opinions about political cases, economic changes, companies and their products.

With the emergence of social media platforms, the sentiment analysis studies are shifted from document level analysis (Bruce and Wiebe, 1999; Wiebe et al., 1999; Wiebe, 2000) towards sentence or phrase level analysis (Morinaga et al., 2002; Yi et al., 2003; Kim and Hovy, 2004; Yu and Hatzi-vassiloglou, 2003; Wilson et al., 2005). Recent years showed that syntactic and/or semantic analy-sis outperforms baseline sentiment analysis meth-ods in many areas such as aspect-based and com-parative opinion mining (Hu and Liu, 2004; Liu, 2012; Balahur et al., 2014). In order to reach this level of analysis, many other natural language pre-processing stages are required; i.e. tokenization, normalization, parts-of-speech tagging etc...

As in all other natural language processing (NLP) problems, the most widely studied lan-guage for sentiment analysis is English. However, studies for morphologically rich languages are not mature yet. Abdul-Mageed et al. (2014) used a supervised, two-stage classification approach em-ploying morphological, dialectal, genre specific features besides basic ones for a morphologically rich language, Arabic. Jang and Shin (2010) pro-poses an approach for agglutinative languages and test their method on Korean short movie reviews and news articles. Wiegand et al. (2010) investi-gate the impact of negation in sentiment analysis of German.

In the literature, it has been shown several times that Turkish, due to its highly inflectional and derivational structure, poses many different prob-lems for different NLP tasks when compared to morphologically poor languages. By this prop-erty, previous NLP research on Turkish language pioneered the studies for many similar languages. On the other hand, sentiment analysis studies for Turkish are very preliminary; although there ex-ist a couple of studies on sentiment classification of movie reviews, political news, fairytales (Vural et al., 2013; Kaya et al., 2013; Boynukalin, 2012; Seker and Al-naami, 2013), there exist very few studies on sentiment analysis of social media posts (Çetin and Amasyalı (2013a; 2013b)).

With the emergence of new tools dealing with automatic language processing of social media texts (Eryi˘git, 2014), it is now becoming possible to integrate them into higher level applications; i.e. sentiment analysis in our case. But, the following issues still reside as open questions:

1. the impacts of each NLP layers on sentiment analysis.
2. information (e.g. stems, main POS tags, in-flectional features) to use from the outputs of beneficial layers.

In this paper, for the first time in the literature, we investigate and report the impact of the prepro-cessing layers (namely, tokenization, normaliza-tion, morphological analysis and disambiguation) on the sentiment analysis of Turkish social media texts. In order to show the maximum sentiment analysis performance to be achieved with flawless NLP tools, we used a hand-annotated sentiment corpus with gold-standard linguistic features.

2 Turkish

Turkish is an agglutinative language where each stem may be inflected by multiple suffixes. Every new suffix concatenation may change the meaning of the word or redefine its syntactic role within the sentence.

This feature of Turkish yields to relatively long words (having higher number of char-acters when compared to other languages). As an ordinary example of this situation, the Turk-ish word "yapabilirmi¸scesine" can be translated as *"as if he/she is able to do"* into English. In addition, the example shows that the same En-glish statement is expressed by a lesser word count (smaller mean sentence length) in the Turkish side. Therefore, semantic analysis of Turkish social me-dia texts is more risky to be defeated by the erroneous writings within this informal domain. The various problems observed in the Turkish Tweets are presented in detail in Toruno˘glu and Ery-i˘git (2014); these are mainly the missing vowels, diacritics, the usage of emoticons, slang words, emo-style writings, spoken accents and high occurrence of spelling errors. The lower word count within a sentence leads to strict dependencies be-tween words in Turkish and the only one single misspelled word can ruin the understandability of the whole sentence. This indicates the importance of normalization preprocessing stage for Turkish differently from English.

POS tagging task for other languages is per-formed in two stages for Turkish: morphological analysis and morphological disambiguation. Mor-phological analysis of a single word can produce several possible analysis regardless of the context in sentence. However, only one of them is correct in its context. The correct analysis can be selected by morphological disambiguation process on the morphological analysis results. Linguistic infor-mation about the word and possible relations with other words in the sentence can be extracted from the correct analysis.

3 The Used Data Set

For this study, we collected a twitter Turkish sen-timent corpus mainly from the telecommunication domain. The data is retrieved from the Tweeter API by querying a predetermined list of keywords. The time frame of the collected data was between May, 10th of 2012 and July, 7th of 2013. We re-fined the corpus from non-Turkish tweets through a language specifier based on a "Language Detec-tion Library for Java"[1]. For the manual annotation of our corpus, we used TURKSENT (Eryi˘git et al., 2013) - a sentiment annotation tool which allows us to annotate the corpus on the following layers: general and target based sentiment, text normal-ization, morphology and syntax. For this study, we used only the general sentiment, the normal-ization and the morphological annotation layers of the tool.

Since the sentiment annotations depend on sub-jective decisions of the human annotators, we ap-plied an inter-annotator agreement filter to in-crease the confidence level of our sentiment anno-tations. Our final dataset consists of 12790 tweets manually normalized, morphologically analyzed and classified between 3 sentiments (3541 posi-tive, 4249 negative and 5000 neutral) agreed by two human annotators.

4 Feature Extraction Methods

In this study, we treat the sentiment detection of a tweets as a multi-class classification problem. We used support vector machines (SVM) in or-der to classify the tweets into one of the three classes (positive, negative, neutral). When we extract unigrams from all collected data without preprocessing and feature filtering, we get 97472 unique features. This amount of features is ex-tremely huge for machine learning algorithms, be-cause more features ends up with more training time and more resources. In addition to time and resource constraints, irrelevant features may also ruin the steady nature of the trained model. Since feature extraction is an indispensable stage of machine learning algorithms, we applied an ex-traction method utilizing *Inverse Document Frequency (IDF).* While *Term Frequency* is easier and sipler than IDF, it is not convenient if there are lots of recurring parts of texts which is the case for our study. Tweets are treated as single documents while calculating the document frequencies in IDF. After the calculation of IDF values of all unigram features, we filter them according to our proposed filtering algorithm Min-ClosestTh given below.

MinClosestTh. A small IDF value indicates a characteristic feature for a given class. But, in order for a feature to be discriminative between different classes, the difference between its IDF values should be bigger than a given threshold. In other words, a feature having similar IDF values for two classes does not help for the discrimination of these classes. For example, a stopword or a keyword which is used to retrieve data from Twitter API will have similar small IDF values for all classes. In the light of these observations, after testing with several feature extraction methods2, we found that MinClosestTh performed the best. In this approach (Equation 1), we find the diffrence between the smallest and the second smallest IDF3 value for a feature among all classes. The features, falling outside of this threshold are removed from the feature set.

$$|minIDF - medianIDF| > threshold \quad (1)$$

Figure 1 shows the histogram of |minIDF - medianIDF| difference distributions. One should notice that a determined threshold value will also determine the number of features to be used in the experiments; all the words entering to the bins greater than the threshold value will be included into the feature set.

[1]It is available on https://code.google.com/p/language-detection/

In order to select a good threshold value for further experiments, we investigate the sentiment analysis performance with different threshold values (0.1, 0.15, 0.3, 0.5 and 0.8). These are given in small line chart in Figure 1. As seen from the figure, the maximum f-measure is achieved at 0.15. F-measure is not the only metric to select the optimum threshold since the total feature count should also be considered. For example, the number of features in the feature set is 18536 when the threshold is chosen 0.1 and 3685 when 0.15. Although the difference between f-measures is not dramatic, the lesser number of features is preferable. We selected 0.15 for further experiments since as seen from the figure, the performance drops consistently without having any important difference in feature counts.

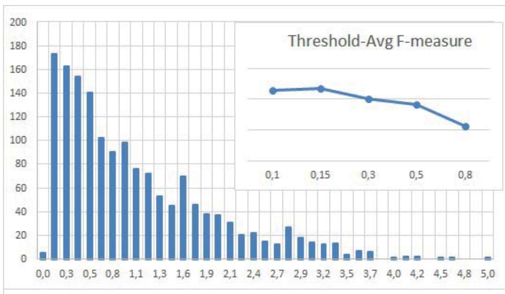

Figure 1: minIDF - medianIDF Histogram and Related Performance

5 Natural Language Preprocessing Layers

Turkish is an agglutinative language and stems can be transformed theoretically to unlimited number of variations with derivational affixes.

2Due to space constraints, we only provide here our best model.

3Since we have only three classes, the second smallest IDF is represented as medianIDF in Equation 1.

Moreover, all these different variations of a word may not make a difference on sentiment classification of tweets. Therefore, we want to polarize features which have the similar impact on sentiment to the same pole, and make explicit the difference between poles. We applied mainly three different NLP preprocessing layers (explained in previous sections) to transform features from original versions to the desired representations. Below we give the information extracted from the output of these layers.

Normalization. We used the normalized forms of the words before extracting the features. For instance, "t¸skkrler" is normalized as "te¸sekkürler" (thanks).

Stemming. Stems of words have more general coverage than surface forms. To match different surface forms of a word into one simple token, we used stemming by deleting all inflectional groups and tags from its correct morphological analysis. For instance, "uzmanlar" (specialists), "uzmanlı˘gı" (his/her specialty), "uzmanlık" (specialty) are derived from the same stem "uzman" (specialist). All three forms are turned into their stem "uzman".

Negation. As stated in Wiegand et al. (2010), the detection of negation needs extra treatment in morphologically rich languages where the negation may be realized within the word with an affixation rather than a separate individual word. The case holds very frequently for Turkish, that's why our motivation in this section is to model the negation for sentiment analysis.

Model#	Model Name	Avg. F-measure	Accuracy	Feature#
1	no_normalization – no_preprocessing	73.38	73.72	78025
2	normalization	78.05	78.28	39788
3	normalization-stem	78.35	78.63	17855
4	normalization-stem-neg	78.83	79.09	18493
5	normalization-stem-neg-adj	77.93	78.27	23613

Table 1: Sentiment Analysis Experiments Results

Negative indicators -such as the inflectional tags at the output of morphological analysis: "+Neg", "+WithoutHavingDoneSo" (like in use of *regard-less of*, or *without stopping*)- have a power to turn meaning of words into opposite. For instance, "çekmiyor" (meaning *"there is no signal"* for the the telco domain) has a morphological analysis such as "çek +Verb+Neg+Prog1+A3sg" where the stem "çek" translated literally as *to pull* into En-glish. If a feature will be extracted from this word we represent it as "çek+Neg". In addition, nega-tion word, "de˘gil" (means to *not* in English), has the same negative effect on preceding words. We put negation tag if a word contains negative indica-tors, or has "de˘gil" as its successor. For instance, "iyi de˘gil" (*not good*) is represented as "iyi+Neg". Furthermore, we added negation tag to the ad-jective if its successor is a negative verb. "Net göremiyorum." (*I can't see clearly.*) transformed to "Net+Neg gör+Neg". When a word achieved double negation tag because of these conditions, we removed all the negation tags belonging to this word. For example, "sessiz de˘gil" (*not silent*-"siz" suffix matches with *less*, like use in *noiseless*.) converted to "ses", not to "ses+Neg +Neg".

Using adjectives. We performed extra effort for adjectives in this research, because of the gen-eral belief that adjectives have a direct impact on sentiment analysis in comparison with other word types. We added adjectives to the feature set with-out exposure them to filtering by feature extrac-tion methods defined previously. Even if we ap-plied any of the other NLP preprocessing methods on adjectives just like any other word types, we also used surface form of adjectives as an addi-tional feature instead of using only preprocessed versions. For example, we represent the adjec-tive "tatsız" (tasteless) with two different features, "tat+Neg" (taste+Neg) and "tatsız".

6 Experiments and Discussions

In all of our experiments, we used SVM with lin-ear kernel. In order to increase the confidence level of sentiment analysis, we applied 10-fold-cross-validation. The results are presented in terms of macro average of all iterations in Table 1.

We tested with 5 different NLP preprocessing models where each of them is the addition of a new processing layer on top of the previous one.

The first line of the table (no_normalization –no_preprocessing) presents our baseline model. This test is performed on the original version of the data set, in other words without applying any preprocessing during the selection of the feature set. The further experiments are evaluated accord-ing to their preceding experiments, and the perfor-mance improvement of the best model is reported with respect to this baseline.

Table 1 shows that the normalization stage (Model #2) contributes to the sentiment analysis, and increases the overall success by about 5 per-centage points. On the other hand, although the addition of the stemming (Model #3) results in a slight improvement on top of Model #2, this im-provement is not proven to be statistically signif-icant according to McNemar's test. Despite this, Model #3 is considered very valuable since the to-tal number of features is almost reduced by 50%(39788→ 17855). As a result, the lesser number of features provide us the ability to train our clas-sifier by using less time and less resources as we mentioned in Section 4. This yields the possibil-ity of adding more valuable training data to our machine learning algorithm, especially for active learning experiments.

Our final two experiments (Model #4 & Model #5) deal with the addition of some morphological features into sentiment analysis (detailed in Sec-tion 5). Although with the addition of negation (Model #4), we observed a slight improvement in the results, this improvement is again not statisti-cally significant where as it also increases the total number of selected features. A similar case holds for Model #5 again with no statistical significance, but this time with a small decrease.

Our final two experiments (Model #4 & Model #5) deal with the addition of some morphological features into sentiment analysis (detailed in Sec-tion 5). Although with the addition of negation (Model #4), we observed a slight improvement in the results, this improvement is again not statisti-cally significant where as it also increases the total number of selected features. A similar case holds for Model #5 again with no statistical significance, but this time with a small decrease.
As the conclusion, in this study, we showed that normalization is an indispensable stage for sen-timent analysis where as stemming is also very valuable for further studies (e.g. active learn-ing). However, our tested model for the addition of morphological information into the system do not seem well-fitted for this domain. Nevertheless, we may not conclude that the morphological informa-tion such as negation has no impact on sentiment analysis. We rather sense that we need to make further research on the inclusion of morphological features such as using them as separate features instead of the approach defined in here (the con-catenation: Stem+Neg).

7 Conclusion and Future Work

Feature extraction methods provide us to decrease training time of classifiers, and also they have a positive impact on sentiment analysis success rate. We achieved higher sentiment analysis suc-cess rate with less number of features. In addi-tion, we showed how the normalization improve the sentiment analysis on Turkish social media posts. With the normalization preprocessing, we increased the success rate of sentiment analysis from 73.38% to 78.05%, which is the 6.36% rela-tive improvement.

By the addition of morpholog-ical features we saw a slight improvement from 78,05% to 78,83% which is not statistically sig-nificant according to McNemar. However, stem-ming, which is the first morphological feature that we applied, is dramatically reduced the number of features as an advantage of ability to train mod-els with more data. For our future studies, we will work on developing automatic NLP tools to make use of morphological information. Thereby, we want to build an environment for further linguis-tic analysis, such as syntax and semantics. We ex-pect to increase sentiment analysis success by such deep analyzes of language.

8 Acknowledgments

closed for blind review

References

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar:

Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1):20–37.

Alexandra Balahur, Rada Mihalcea, and Andrés Mon-oyo. 2014. Computational

tapproaches to subjec-tivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1):1–6.

Zeynep Boynukalin. 2012. Emotion analysis of Turk-ish texts by using machine learning methods. Ms.

Rebecca F Bruce and Janyce M Wiebe. 1999. Recog-nizing subjectivity:

a case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

Mahmut Çetin and M Fatih Amasyali.

2013a. Ac-tive learning for Turkish sentiment analysis. In *In-novations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on*, pages 1–4. IEEE.

Gül¸sen Eryi˘git, Fatih Samet Çetin, Meltem Yanik, Tanel Temel, and Ilyas Çiçekli.

2013. Turksent: A sentiment annotation tool for social media. In *Pro-ceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 131–134, Sofia, Bulgaria, August. Association for Computa-tional Linguistics.

Gül¸sen Eryi˘git. 2014. ITU Turkish NLP web service.

In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Associa-tion for Computational Linguistics (EACL)*, Gothen-burg, Sweden, April. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004.

Mining and summa-rizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowl-edge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Hayeon Jang and Hyopil Shin. 2010.

specific sentiment analysis in morphologically rich languages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 498–506, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mesut Kaya, Guven Fidan, and I Hakkı Toroslu. 2013.

Transfer learning using twitter data for improving sentiment classification of Turkish political news. In *Information Sciences and Systems 2013*, pages 139–148. Springer.

Soo-Min Kim and Eduard Hovy. 2004.

Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Bing Liu. 2012.

Sentiment analysis and opinion min-ing. *Synthesis Lectures on Human Language Tech-nologies*, 5(1):1–167.

Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 2002.

Mining prod-uct reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349. ACM.

Sadi Evren Seker and Khaled Al-naami. 2013.

Sentimental analysis on Turkish blogs via ensemble clas-sifier. In *Proceedings Of The 2013 International Conference On Data Mining*. DMIN.

Dilara Torunoˇglu and Gül¸sen Eryiˇgit. 2014.

A Clscaded approach for social media text normalization of Turkish. In *5th Workshop on Language Analy-sis for Social Media (LASM) at EACL*, Gothenburg, Sweden, April. Association for Computational Lin-guistics.

A Gural Vural, B Barla Cambazoglu, Pinar Senkul, and Z Ozge Tokgoz. 2013.

A framework for sentiment analysis in Turkish: Application to polarity detec-tion of movie reviews in Turkish. In *Computer and Information Sciences III*, pages 437–445. Springer.

Janyce Wiebe. 2000.

from corpora. In *AAAI/IAAI*, pages 735–740.

Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. 1999.

Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the As-sociation for Computational Linguistics on Compu-tational Linguistics*, pages 246–253. Association for Computational Linguistics.

Michael Wiegand, Alexandra Balahur, Benjamin Roth,Dietrich Klakow, and Andrés Montoyo. 2010.

A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 60–68, Stroudsburg, PA, USA. As-sociation for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005.

Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Con-ference on Human Language Technology and Em-pirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003

. Sentiment analyzer: Extract-ing sentiments about a given topic using natural lan-guage processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE.

Hong Yu and Vasileios Hatzivassiloglou. 2003.

To-wards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Lin-guistics.