Natural Language Engineering **1** (1): 1–**??**. Printed in the United Kingdom

Sentiment Analysis in Turkish: Towards a Complete Framework

Rahim Dehkharghani, Berrin Yanikoglu, Yucel Saygin
{*rdehkharghani,berrin,ysaygin*}@*sabanciuniv.edu*
Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey

Kemal Oflazer
ko@cs.cmu.edu
Carnegie Mellon University – Qatar, Doha, Qatar

(*Received 5 June 2015*)

Abstract

Sentiment analysis has attracted a lot of research interest in recent years, especially in the context of social media. While most of this research has focused on English, there is ample data and interest in the topic for many other languages, as well. In this article we propose a comprehensive sentiment analysis system for Turkish. Our contributions include addressing linguistic issues such as negation and intensification, as well as covering different levels of sentiment analysis such as aspect, sentence, and document levels. We evaluated our methodology on Turkish movie reviews and obtained accuracies ranging from 60% to 79% in ternary and binary classification tasks at different levels of analysis.

1 Introduction

Sentiment analysis and opinion mining aims to extract the embedded polarity from textual data. Sentiment analysis has attracted significant attention in recent years as social media has become a venue for expressing user and customer opinions about products, services, political platforms, etc., providing an opportunity for companies and organizations to gather feedback from a large number of people.

Sentiment analysis is often applied in a language-dependent (e.g., English or Turkish) and a domain-dependent (e.g., movies or hotels) setting. For instance, the term *"big"* is positive for *room size* in hotel reviews but negative for *battery size* in camera reviews. Similar examples may be given for the same concept being objective versus polar accross different languages (e.g. the word "God"). while some of the issues are shared across languages and domains (e.g. the need to handle negation), sentiment analysis systems need to use resources in the language and domain of interest.

English has the richest set of sentiment analysis resources such as SentiWordNet (Esuli and Sebastiani 2006), and SenticNet (Cambria, Olsher and Rajagopal 2014). However, as social media proliferates in many other places where different languages are spoken, it brings together a strong demand for sentiment analysis in those languages.

We focused on Turkish sentiment analysis, due to significant proliferation in the number of users as compared to the overall population. The few works on Turkish sentiment analysis so far have focused on a binary (positive and negative) classification at the document level. More research is needed in order to address different levels of granularity (e.g. sentiment analysis of tweets) and support ternary classification handling neutral/objective opinions.

In earlier work, we have built polarity resources for Turkish such as SentiTurkNet (Dehkharghani, Saygin, Yanikoglu, and Oflazer 2014), and a relatively large polar word set consisting of 2000 words (Dehkharghani *et al.* 2014). In this work, we propose a sentiment analysis system for Turkish and apply it to Turkish movie reviews.[1] The proposed methodology can be employed for other languages with minor changes. Our method works at aspect, sentence, and document levels. Our contributions can be summarized as follows:

2 Related Work

There are few work on sentiment analysis of Turkish texts. Yıldırım, Çetin, Eryiğit and Temel (2015) accomplished a sentiment analysis task on Turkish tweets in the telecommunication domain. They applied a multi-class ternary (positive, negative, and neutral) classification by support vector machines on tweets using features such as inverse document frequency, unigrams and adjectives. They also benefit from NLP techniques such as Normalization, stemming and negation handling. The best accuracy in classifying tweets as three classes in reported as 79%.

Vural, Cambazoğlu, Şenkul and Tokgöz (2012) present a system for unsupervised sentiment analysis in Turkish text documents by customizing *SentiStrength* (Thelwall and Paltoglou 2012) by translating its polarity lexicon to Turkish. SentiStrength is a sentiment analysis tool for English that assigns a positive and a negative score to a segment of text . Authors report 76% accuracy in classifying Turkish movie reviews as positive and negative, using this method.

Kaya, Fidan and Toroslu (2012) investigate sentiment analysis of Turkish political news in online media. Authors use four different classifiers–Naive Bayes, Maximum Entropy, SVM, and the character based n-gram language models– with a variety of text features: frequency of polar word unigrams, bigrams, root words, adjectives and effective (polar) words. They conclude that the Maximum Entropy and the n-gram language models are more effective when compared to the SVM and Naive Bayes classifiers, in classifying Turkish political news. They report an accuracy of 76% in binary classification of political news.

Boynukalın (2012) has worked on emotion analysis of Turkish texts by using machine learning methods. She has investigated four types of emotions: joy, sadness, fear, and anger on a dataset that she built for this purpose. She reports an accuracy of 78% in classifying documents into these four emotions.

[1] These reviews are collected from www.beyazperde.com

Eroğul (2012) explores the use of surface linguistic features such as part-of-speech tags, word unigrams and bigrams, and negation markers. This work relies on a morphological analyser for Turkish, called Zemberek (Akın and Akın 2007)), and reports an accuracy of 85% on classifying Turkish movie reviews (at the document level only) as positive and negative. Note that while many researchers use movie reviews, the exact datasets used to assess different methods are different, making it impossible to directly compare results.

In spite of the existing works on sentiment analysis of the Turkish language, none of the approaches are comprehensive in terms of its coverage of issues and evaluation of its results. In this paper we propose and evaluate a rather comprehensive sentiment analysis system for Turkish including different levels such as sentence, aspect, and document levels incorporating a large variety of surface and deeper linguistic features.

3 Levels in Sentiment Analysis

The most common level of sentiment analysis is done at the document level, in which a polarity label (positive, negative, or neutral) is estimated for the whole document. Document level analysis may lead to information loss in longer documents with mixed sentiment. For example in movie reviews, if an aspect (e.g., action) is positive but another aspect (e.g., director) is negative, the sentiment analyser may classify this document as neutral while in fact it has mixed sentiment. Finer grain analysis is required to address this issue, as indicated below:

- *Word level:* Assigning a sentiment polarity to a word is not very easy, as a word may have different polarities in different domains or even in the same domain (see the "big" example given earlier).
- *Phrase level:* A phrase is an ordered (not necessarily consecutive) list of n terms within a sentence and a sentence is composed of one or more phrases, possibly with different sentiments. For example the sentence below has two phrases with two different sentiments (one is shown in italic):

  Ben beğendim, *ama herkes beğenmedi.*
  (I liked it, *but not everyone did.*)
- *Sentence level:* Sentiment analysis of sentences may follow word and phrase-level analyses. If the sentence has a mixed polarity (both positive and negative due to multiple aspects or phrases), one can assign an overall polarity based on relative sentiment strengths of the components/phrases inside the sentence.
- *Aspect level:* Aspects are different perspectives relating to the review item, e.g., "room" in hotel reviews or "plot" in movie reviews. Each sentence in a domain may include several aspects and the polarity of each aspect may be different from the overall polarity of the sentence. For example the sentence below has two phrases about two separate aspects, one with positive and the other with negative sentiment (one is shown in italic):

  oyunculuk iyi, *ama efektleri sevmedim*
  (the acting is good, *but I did not like the effects*)

- *Document level:* This is the coarsest level and attempts to estimate the overall polarity of a document. Often document polarity is aggregated from the estimated polarity of the constituent words or sentences. Previous work (Meena and Prabhakar 2007) (Gezici, Yanikoglu, Tapucu and Saygın 2012) have shown that sentence-level analysis is effective and initial and last sentences may have higher influence on document polarity, compared to sentences in the middle.

4 Natural Language Processing Issues in Sentiment Analysis

An effective sentiment analysis system must handle various linguistic markers such as negations, intensifications, and conditional constructions, in order to make more precise sentiment classifications. Most of these marker are language-specific and their extraction requires language-specific tools (e.g., morphological analyzers and parsers), while some others such as emoticons are considered language-independent.

Below we group the issues that we rely on for Turkish sentiment analysis, into two subsets: "linguistic" and "other" issues. Here, we present only the challenges, while proposed solutions are presented in Section 5.3.

4.1 Linguistics issues

- *Negation:* Negation markers can switch the polarity of a predication or main verb in their scope. The following sentence is a simple negation form by using the predication negation marker "değil" (is/am/are not):

  ...20 defa izlemişimdir, *pişman değilim.*
  (...probably watched it 20 times, *I am not regretful.*)

  where we have a negative to positive change in the sentiment as "pişman" (regretful) is negated by "değilim" (I am not).
  The second example provides a more complicated negation form by two negated verbs where the underlined morphemes in words mark negation:

  sevmedim diyen *çıkmadı*
  (*no one came out* saying that they *did not like it*)

  where polarity first switches to negative with " sevmedim" (I did not like) and back to positive within "çıkmadı" (no one came out).
- *Intensification:* Intensifiers such as "çok" (very) and "biraz' (so so/a little) modulate the polarity of a term stronger or weaker. For example, the adjective "iyi" (good) is strengthened in "*çok* iyi" (*very good*) or weakened in "*biraz* iyi" (*so so good*).
- *Conditional sentences:* These sentences may change the *apparent* polarity of a sentence. For example the sentence below indicates a less positive sentiment than what is indicated by the existence of a high score of 10.

  Çok uzun olmasaydı, 10 verirdim.
  (*If* it was not too long, I would have given it a 10.)
- *Rhetorical questions:* The polarity of these sentences usually differ from what

appears on the surface–that is the expression is formally a question sentence but is not used to elicit an answer; it rather is used to convey a variety of sentiments. For example, in the sentence below, the overall sentiment is made positive with the addition of the question suffix "mi", while "sevmez" (does not love) has negative polarity.

İnsan bu filmi sevmeyebilir *mi*?
(*Can* one not like this movie?)

- *Sarcastic phrases:* Sarcasm detection may be the most challenging issue in language processing tasks. This task has obtained very low accuracy (52%) even in English (Liu 2012). A sarcastic statement such as "harika bir film olmuş!" (it was a great movie!) can only be detected by the disagreement with it and the whole of the (negative) review and slightly hinted by the exclamation mark.
- *Idiomatic uses:* An idiom is a combination of words whose meaning is a compositional combination of the meanings of its constituent words. The challenging issue in idioms is that the polarity of an idiom cannot always be extracted automatically by using the polarity of terms included within the idiom. For example a commonly used idiomatic compound verb in Turkish is "göz boyamak" (*to deceive* – literally to paint the eyes) which has a negative sentiment while its constitutents "göz" (*eye*) and "boyamak" (*to paint*) are neutral terms when considered separately.

4.2 Other issues

We grouped here issues that are not language-dependent but represent issues or techniques that may need to be addressed by sentiment analysis systems across different languages.

- *Emoticons:* Emoticons can help estimate the polarity of a sentence. Normally positive emoticons (e.g. ":)") appear in positive sentences and negative ones (e.g. ":(") appear in negative sentences. As the number of emoticons used in a language increases, they start to carry more and more of the sentiment.
- *Conjunctions:* Conjunctions can help estimate the polarity of the two terms around the conjunct, with the help of one another. For example two adjectives conjoined by "ama" (but) are supposed to have opposite polarities, while they often have the same polarity when they are conjoined by "ve" (and). This observation was made and used to estimate word-level polarities in previous work (Hatzivassiloglou and McKeown 1997).
- *Domain-specific indicative keywords*: The polarity of sentiment keywords can change across domains. Furthermore, each domain has some keywords that are good clues for estimating the polarity of a sentence/review that includes those keywords. For example the phrase "kaçırmayın" (do not miss it) at the end of a movie review is a commonly used positive phrase in the movie domain.
- *Background knowledge:* Sentiment analysis systems require background

knowledge for classifying special kinds of sentences such as: *"of those rare films that makes me feel that I am present in the film"*. In this sentence, the key issue is that the feeling of being present in the film is a positive emotion, which is the background knowledge necessary to understand the sentiment. It is however extremely difficult with the current state of the art in natural language processing to extract such information.

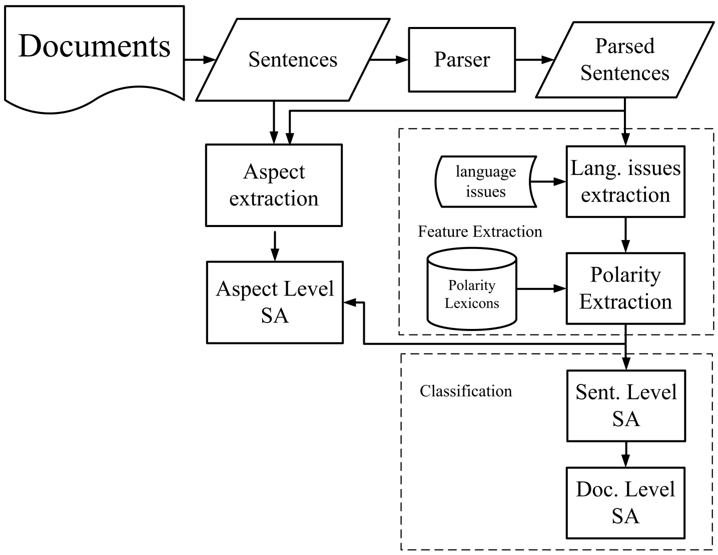

Fig. 1. The proposed system as a flowchart

5 Our Methodology

In this section, we first present an overview of our system and then elaborate on each component of the system, explained in the following subsections. The system consists of several components as illustrated in Figure 1. The input is a document (a movie review) which is segmented into sentences and then each sentence is fed to a parser (Eryiğit 2014) that provides the dependency tree structure of the sentence and morphological analysis for each word. This structure is used in aspect-level polarity classification (see Section 5.2.3).

We assign polarity scores to word n-grams (unigram, bigram, and trigram) by using the following polarity lexicons: SentiTurkNet, our own polar word list, and translation of the SenticNet (see Sections 5.1 and 5.2.4).

After assigning polarity values to terms in a sentence and covering linguistic and other related issues, we do a sentence level polarity classification first. A document level sentiment classification is then accomplished by using features listed in Table 6 with four additional features (compared to Table 5) indicating the estimated polarities of the first and last sentences in the document.

Table 1. **Parse tree generated by using the ITU parser for the sentence "hoş vakit geçirmek için seyredilebilir" (it can be viewed for an enjoyable time).**

1	Bence	ben	pron	pers	A1sg.pnon.equ	0	root
2	hoş	hoş	adv	adv	–	3	modifier
3	vakit	vakit	noun	noun	A3sg.pnon.nom	4	object
4	–	geçir	verb	verb	Pos	5	deriv
5	geçirmek	–	noun	Inf1	A3sg.pnon.nom	6	object
6	için	için	postp	pcnom	–	0	modifier
7	seyredilebilir	seyredil	verb	Able	Pos.aor.a3sg	0	root

5.1 NLP Tools and Polarity resources

We rely on a parser and three polarity lexicons in this work.

- *ITU Turkish Parser (Eryiğit 2014):* This parser receives a Turkish sentence as input and builds a dependency tree with morphological analyses for every token in the sentence. The output of this parser for the sentence "bence hoş vakit geçirmek için seyredilebilir." (It can be viewed for an enjoyable time) is illustrated in Table 1.
- *Polar word Set (PWS):* We have semi-automatically generated a list of polar Turkish terms including 1000 positive and 1000 negative terms using the method proposed by Hu and Liu (2004). This method uses synonymy and antonymy relations between terms to generate a large polar word set starting from a small seed set.
- *SentiTurkNet (*STN*):* We have developed the first Turkish polarity resource based on the Turkish WordNet (Bilgin, Çetinoğlu and Oflazer 2004), where three polarity scores are assigned to each Turkish synset (set of synonyms) indicating its positivity, negativity, and neutrality levels. This resource consists of about 15,000 synsets and 1.47 terms per synset in average.
- *SenticNet (*SN*):* This resource assigns different value to each term as its *pleasantness, attention, sensitivity, aptitude,* score, as well as an *overall-polarity.* Each of these features has a value between -1 and $+1$ as the most negative and the most positive polarities respectively. We translated this resource to Turkish by a bilingual dictionary named *SesliSözlük* (Seslisözlük Group 2014) and used only the *overall polarity* of each term (or phrase) as the sentiment polarity. This lexicon contains about 14,000 entries (words and phrases).

The above mentioned resources cover different sets of terms and may assign different polarity scores to the same term. Also they assign polarity scores (real values) or labels to each term, making a comparison of these resources difficult. The word "güzel" (good/beautiful), for example, is labelled as positive in *PWS*; has a score of *+0.44* in *SN*; and [pos, neg, obj] scores of [1, 0, 0] in *STN*. Although we believe that using Turkish polarity lexicons are more effective than using the translation of English resources, we used both methods to compare their effectiveness in the classification of Turkish movie reviews.

5.2 Sentiment Analysis for Turkish at Different Levels

Our system is designed to address different levels of sentiment analysis: words, phrases, sentences, aspects, and overall document, as explained below.

5.2.1 Word level

We extract the polarity of a given word using the polarity lexicons described in Section 5.1 (*PWS*, *SN* and *STN*). Polarities of constituent words are used in all subsequent steps, including feature extraction of sentence and document-level analysis.

Since *STN* is derived from the Turkish WordNet, different senses of a word may be associated with different polarity scores. The solution in this case is to do Word-Sense Disambiguation (*WSD*) to find the correct sense of the word considering the context; however, *WSD* is an ongoing problem in both English and Turkish and is out of the scope of this work. In the proposed system, we narrow the possible senses of a word by relying on the morphological features–mostly the Part Of Speech (*POS*). Exploiting the *POS* tags for this purpose improves the polarity extraction, compared to randomly choosing the word-sense in a context. In *SN*, we use only the *overall-polarity* score of each word or phrase (sequence of words). In *PWS*, only the polarity label (positive or negative) is available, which indicates the overall polarity of words.

Word-level polarities found here are then combined considering linguistic markers mentioned in Section 4 and modified by the methods proposed in Section 5.3. The modified polarity scores/labels are used as word polarities in all subsequent steps (phrase, sentence, aspect and document level classifications).

5.2.2 Phrase level

We use the dependency parse tree produced by the parser described above to identify disambiguated phrase structures. We generate structures with any number of terms, for example if term t_i is related to (dependent on) term t_j, t_j is related to t_k, and t_k is related to t_l, the phrase "$t_i t_j t_k t_l$" is extracted from the sentence. The relations let us focus on the main predications or relevant modifications or conjunctions in the sentence, ignoring words that may not be relevant for sentiment analysis. Looking up the words in the resource we have built, provide initial estimates of sentiment.

We do not explicitly do sentiment analysis in the phrase level; instead, we use the output–extracted phrases by dependency parse tree–in the aspect level sentiment analysis.

5.2.3 Aspect level

We compiled a list of aspects (A) in movie domain and proposed a novel method for estimating the polarity of each aspect. After identifying an aspect a_j in a sentence S, we identify those relations to encode basic predications. An example sentence is given below.

Oyunculuk iyi, *ama efektleri pek sevmedim.*
(The acting is good, *but I did not like the effects that much.*)

In this sentence, the two phrases "oyunculuk iyi" (the acting is good) and "efektleri sevmedim" (I did not like the effects) are extracted from the dependency tree ignoring other words that do not necessarily have much effect on the sentiment. We then compute the average polarity (positivity and negativity) of all such relations involving the aspect a_j in sentence S by means of two terms $P(a_j)$ and $N(a_j)$ that indicate the average positivity and negativity scores of aspect a_j, using Equations (1) and (2).

$$P(a_j) = \sum_{\forall n_k \in NG, s.t.\ a_j \in n_k} \frac{\sum_{t_i \in n_k} pos(t_i)}{|n_k|} \tag{1}$$

$$N(a_j) = \sum_{\forall n_k \in NG, s.t.\ a_j \in n_k} \frac{\sum_{t_i \in n_k} neg(t_i)}{|n_k|} \tag{2}$$

where NG is the set of all relational structures generated by the dependency parse tree; and n_k is a relational structure in the sentence; $|n_k|$ is the number of tokens in n_k; and $pos(t_i)$ and $neg(t_i)$ are positivity and negativity scores of term t_i, as extracted from STN.

These relational structures consist of two, three, or more words that are structurally related together in the dependency parse tree. In these equations, if $P(a_j) > N(a_j)$, a_j is classified as positive, if $P(a_j) < N(a_j)$, a_j is classified as negative, or neutral otherwise. The list of aspects is provided in Table 2.

5.2.4 Sentence level

We start sentence level sentiment analysis by automatically segmenting each document to its sentences by using punctuation, capitalization, and emoticons. Then, we extract 16 features given in Table 5 from each sentence to be used in classification task. The classifier is trained with 2,700 labelled (as pos, neg, or obj) sentences in the Turkish movie reviews and evaluated by 5-fold cross validation.

5.2.5 Document level

We address the document level sentiment analysis similar to the sentence level analysis, using 20 features given in Table 6. The classifier is trained by 1000 feature vectors which have been extracted from 1000 labelled documents (as pos, neg, or obj) in the Turkish movie reviews. We also benefit from additional four features

($f_{17} - f_{20}$) for this level to highlight the effect of the first and last sentences in the document. The evaluation method for this classifier is again 5-fold cross validation.

5.3 Handling Linguistic Issues in Turkish

In this section, we propose our solutions for most of the linguistic issues discussed in Section 4 and leave some of them as future work. We also address some additional relevant issues.

The proposed methods are applied on words and sentences to change their polarity if applicable. The initial polarity scores and labels for words are obtained from three polarity lexicons as explained in Section 5.1; these are changed if necessary through the handling of linguistic and other issues.

5.3.1 Linguistic issues

- *Negation:* We covered different kinds of negation in Turkish and were able increase the classification accuracy by about two percentage points.
  - — The predication negation marker “değil” (is/am/are not) switches the sentiment of the preceding the words. For example in the sentence “ama kötü bir film de değil” (but it is not a bad movie either) the marker “değil” switches the negative polarity of “kötü” (bad) to mostly positive.
  - — Morphemes “ma” and “me” in verbs negate the polarity of a verb. For example “sevdim” (I liked) has positive sentiment but sentiment changes to negative when the morphological negation is introduced with the morpheme “me” in “sevmedim” (I did not like). For this, we rely on the disambiguated morphological representation of the verbs provided by the dependency parser.
  - — Morphemes“lu” and “suz” derive adjectives from noun with the semantics of “with” or “without” respectively. For example the noun “kusur” (fault) is a negative term and morphemes “lu” and “suz” generate adjectives “kusurlu” (faulty) and “kusursuz” (flawless) which have negative and positive sentiments respectively.

  An erroneous negated case in our system is “Film güzel, değil mi?” (The movie is good, isn’t it?), which has received a negative sentiment polarity. The reason is the inability of the system in understanding that “değil” (is not) is a part of phrase “değil mi” (isn’t it), and not a part of “Film güzel değil” (the movie is not good). Note that without the comma, the sentence “Film güzel değil mi?” would be interpreted as a question asking whether the movie is not good.
- *Intensification:* We compiled a set of intensifiers in Turkish listed in Table 3. For strengthening intensifiers we double the sentiment value and for weakening intensifiers we halve it. This has contributed about a percentage points to our classification accuracy.
- *Conditional sentences:* We cover this only by adding a boolean feature to the classification features (Tables 5 and 6) indicating the conditionality of

a sentence. In other words, we get help from the classifier to estimate the polarity of conditional sentences based on observed conditional sentences in the training set. This issue needs further investigation that we have left for future work.
An erroneous sample is “beğenmeseydim yorum yapmazdım” (if I did not like [the movie], I would not write a comment). This sentence is misclassified as negative because of the phrase *I did not like*.

- *Rhetorical questions:* We attempt to cover this issue by adding a boolean feature to the classification task, which indicates if a sentence is interrogative. However, capturing only the rhetorical questions (not all interrogative sentences) needs further investigation that we have left as future work.
  An erroneous sample for this issue is “Titanik gibi bir film nasıl sevilmeyebilir?” (How a movie like Titanik can be disliked?), in which the classifier is unable to understand the embedded positive sentiment in the sentence.

Table 2. **The list of chosen aspects from Movie domain for our system.**

aksiyon (action), oyuncu/aktor (actor), müzik (music), sahne (scene), efekt (effect), senaryo (scenario), oskar (oscar), yönetmen (director), animasyon (animation)

Table 3. **A subset of strengthening and weakening intensifiers.**

Strengthening (very/really): *baya(ğı)*, *gayet*, *çok gerçekten*, *iyice*, *cidden*
Weakening (a little/almost): *biraz*, *azcık*, *yaklaşık*

Table 4. **A subset of domain-specific indicative terms/phrases in Turkish movie reviews.**

izleyin (watch it), iyi seyirler (happy viewing), *izlemeli*, *izlemek gerek* (should be watched), *kaçırmayın* (do not miss it), *izlenebilir* (could be watched)

5.3.2 Other issues

In this work, we covered only three issues from Section 4.2:

- *Emoticons:* We compiled a list of 50 positive and 50 negative emoticons and marked their presence with appropriate features.

- *Domain-specific indicative keywords:* We gathered a list of 20 keywords and key phrases that indicate positive sentiment in Turkish movie reviews. A subset of these keywords and keyphrases is listed in Table 4. Again we mark their presence with appropriate features.
- *Conjunctions:* We apply the idea proposed by Hatzivassiloglou and McKeown (1997) to Turkish, by using the conjunctions "ama/fakat" (but) and "ve" (and). Two adjectives conjoined by "and" are supposed to have the same polarity while they will have most probably the opposite polarity when conjoined by "but". Two examples from Turkish movie reviews are given below.

  Film güzel *ama* çok uzun.
  (The film is good *but* too long.)
  Film güzel *ve* heyecanlı.
  (The film is good *and* exciting.)

  In the former example, our approach estimates the polarity of "çok uzun" (very long) as negative because it already knows that "güzel" (beautiful/good) is positive.
  An erroneous sample for this issue is given below:

  Konu *çok basit ama kötü* de değil
  (The theme is *too simple but also not bad*)

  In this case, although two adjectives with negative polarity seem to be conjoined by *but*, they actually belong to different phrases. A comma after *basit* could clarify this sentence.
  Conjoined adjectives (although rare) help to increase the classification accuracy only about 0.5 percentage points.

5.4 Features for sentence and document classification

The 16 and 20 features used in sentiment classification of sentences and documents are listed in Tables 5 and 6, respectively. Features $f_1 - f_{16}$ in two tables are similar (just the input is either a sentence or a document), but $f_{17} - f_{20}$ are additional features used only in the document level.

Below we explain the features in some detail. The term "review" used for feature explanations irefers to a sentence or document, in sentence and document level sentiment analyses respectively.

- $f_1 - f_4$: The first four features capture the average polarity of terms in a review, computed using two separate resources that assign numerical polarity scores to each term. In *SN*, we label a term as positive if its polarity score is non-negative, otherwise it is considered negative. In *STN*, three polarity scores are assigned to each Turkish synset but we use only positivity and negativity values as features, as the neutrality score depends on these two scores.
- $f_5 - f_6$: These features indicate the number of positive and negative polar terms in each review, as computed according to the *PWS*.
- $f_7 - f_8$: These features indicate the number of positive and negative emoticons in the review.

Table 5. **Features used in sentiment analysis of a sentence, S. SN, PWS, and STN respectively stand for SenticNet, PolarWordSet, and Senti-TurkNet.**

f_1: *average positive score of words in S using STN*
f_2: *average negative score of words in S using STN*
f_3: *average score of positive words in S using SN*
f_4: *average score of negative words in S using SN*
f_5: *number of positive words in S using PWS*
f_6: *number of negative words in S using PWS*
f_7: *occurrence of positive emoticons in S*
f_8: *occurrence of negative emoticons in S*
f_9: *number of adjectives and adverbs in S*
f_{10}: *number of (first letter) capitalized words in S*
f_{11}: *number of domain-specific indicative words in S*
f_{12}: *length of sentence (number of tokens in S)*
f_{13}: *is S a conditional sentence?*
f_{14}: *is S an interrogative sentence?*
f_{15}: *is S a negated sentence?*
f_{16}: *is S an exclamative sentence?*

- $f_9 - f_{12}$: These simple features are defined based on our three assumptions: (1) the higher the number of adjectives and adverbs in a review, the higher the chances of its subjectivity; (2) the higher the number of initial capital words in a review, the greater the chances of neutrality of the review (capitalized terms are proper nouns which are generally neutral); and (3) the higher the number of domain-specific indicative terms in a review, the greater the chances of positivity for the review.
- $f_{13} - f_{16}$: These features capture the interrogative, conditional, negated, or exclamative form of a sentence. These features can be extracted from the output of the parser.
- $f_{17} - f_{20}$: These polarities of the first and last sentences in the document are used as features for document level sentiment analysis, following the sentence level analysis. Generally the first and last sentences are more subjective than the middle sentences because many people write their ideas more clearly in the first and last sentences.
  We analysed the relationship between the document polarity and the polarity of its first and last sentences. Table 7 shows the conditional probabilities of the document polarity given the sentence polarity. For instance 76% of documents with positive sentiment have a positive first sentence. As also shown in previous work, these numbers also indicate that the first sentence polarity is especially indicative of document polarity.

Table 6. **Features used in sentiment analysis of a document, D. SN, PWS, and STN respectively stand for SenticNet, PolarWordSet, and SentiTurkNet.**

f_1: *average positive score of words in D using STN*
f_2: *average negative score of words in D using STN*
f_3: *average score of positive words in D using SN*
f_4: *average score of negative words in D using SN*
f_5: *number of positive words in D using PWS*
f_6: *number of negative words in D using PWS*
f_7: *occurrence of pos. emoticons in D*
f_8: *occurrence of neg. emoticons in D*
f_9: *number of adjectives and adverbs in D*
f_{10}: *number of (first letter) capitalized words in D*
f_{11}: *number of domain-specific indicative words in D*
f_{12}: *length of document (number of tokens in D)*
f_{13}: *Does D contain a conditional sentence?*
f_{14}: *Does D contain an interrogative sentence?*
f_{15}: *Does D contain a negated sentence?*
f_{16}: *Does D contain an exclamative sentence?*
f_{17}: *avg. positive score of words in first sentence of D*
f_{18}: *avg. negative score of words in first sentence of D*
f_{19}: *avg. positive score of words in last sentence of D*
f_{20}: *avg. negative score of words in last sentence of D*

5.5 Classifier Training

For sentence classification, 16 features in Table 5 are used with a Logistic Regression (LR) classifier (Hosmer and Lemeshow 2012). The evaluation is done using 5-fold cross validation over training data of 2700 sentences. Both binary and ternary classifiers are trained separately at this level.

For document classification, we use the 20 features in Table 6. The classifier and evaluation methods are the same as in sentence level analysis, using logistic regression classifiers and 5-fold cross-validation for evaluation.

6 Experimental Evaluation

We evaluated the proposed approach in terms of its accuracy of classifying sentences, documents and aspects, in both binary and ternary classification scenarios, using 5-fold cross-validation on training data.

Table 7. **Conditional probability of the document polarity given the polarity of the first or last sentence.**

Document	First sentence		
	positive	negative	neutral
positive	0.76	0.01	0.23
negative	0.01	0.79	0.20
neutral	0.13	0.04	0.83
Document	**Last sentence**		
	positive	negative	neutral
positive	0.76	0.05	0.19
negative	0.03	0.56	0.41
neutral	0.13	0.10	0.77

6.1 Dataset

We used a subset of Turkish movie reviews as dataset and manually labelled 1,000 randomly chosen documents from the dataset as positive, negative, or neutral.We also labelled 2,700 sentences appearing in these documents as positive, negative, or neutral. The distribution of [positive, neutral, and negative] sentences and documents are close: [50%, 30%, 20%] and [52%, 29%, 19%] respectively [2]. Finally, we also manually labelled all appeared aspects in the above mentioned sentences, which resulted in about 2,000 aspect mentions labelled as positive, negative or neutral.

We did not include the label "mixed" in our labelling; instead we chose the dominant sentiment in a mixed review and labelled it accordingly.

6.2 Dealing with unbalanced data

As mentioned above, our dataset is unbalanced in favour of positive reviews, which causes biased results for positive samples (sentences and documents) during the classification. To avoid this problem, we balanced the dataset by re-sampling under-represented classes. This technique increased per-class classification accuracies (Tables 10, and 11), while the overall accuracy over all classes did not change much.

6.3 Results

The accuracies obtained from binary and ternary classifications on sentence and document levels are presented in Tables 8 and 9. Using all the features, we ob-

[2] This subset is available from the first author webpage at http://myweb.sabanciuniv.edu/rdehkharghani/sentiment-analysis-in-turkish/

tained 73.42% and 79.06% accuracies in binary sentence and document classification problems, respectively. For ternary classification, results are 60.33% and 73.01% for sentence and document levels. As expected, higher accuracies are achieved at document level (due to larger context) and binary classification problems (simpler problem).

Table 8. **Sentence level binary and ternary classification accuracy (%) by Logistic Regression using 5-fold Cross Validation .**

Feature Subset	Binary	Ternary
f_1-f_2	59.73	59.33
f_3-f_4	59.00	58.74
f_5-f_6	63.24	59.61
f_7-f_8	51.79	49.20
f_9-f_{12}	51.50	59.20
f_{13}-f_{16}	57.99	59.07
f_1-f_4	59.73	60.00
f_1-f_6	70.05	60.12
f_1-f_8	70.40	60.08
f_1-f_{12}	72.28	60.14
all : f_1-f_{16}	**73.42**	**60.33**

Table 9. **Document level binary and ternary classification accuracy (%) by Logistic Regression using 5-fold Cross Validation.**

Feature Subset	Binary	Ternary
f_1-f_2	75.04	69.30
f_3-f_4	76.57	70.70
f_5-f_6	75.68	70.61
f_7-f_8	51.01	48.42
f_9-f_{12}	74.15	69.12
f_{13}-f_{16}	73.50	69.10
f_{17}-f_{20}	78.02	72.30
f_1-f_4	77.44	71.10
f_1-f_6	77.50	71.22
f_1-f_8	78.25	71.20
f_1-f_{12}	78.42	71.34
f_1-f_{16}	78.64	71.51
all : f_1-f_{20}	**79.06**	**73.01**

We also performed an aspect-based sentiment analysis and achieved *70%* and *79%* accuracies in ternary and binary classifications, respectively.

Considering a simple classification system which uses only the positivity and negativity scores of words that would correspond to features $f_1 - f_2$ as a baseline, we could increase the classification accuracy over the baseline by about 4 percentage points, at document level (75.04 vs 79.06 and 69.30 vs 73.01%).

The confusion matrix for both binary and ternary classifications are given in Tables 10 and 11. Each value in these tables shows the per-class accuracy (diagonal values in matrix), separately for positive, negative, and neutral classes in ternary classification and for positive and negative classes in binary classification.

Table 10. **Confusion matrix for binary classification of sentences and documents.**

Document level		
True/Estimated	positive	negative
positive	0.86	0.14
negative	0.27	0.73
Sentence level		
True/Estimated	positive	negative
positive	0.92	0.08
negative	0.67	0.33

Table 11. **Confusion matrix for ternary classification of sentences and documents.**

Document level			
True/Estimated	positive	negative	neutral
positive	0.67	0.20	0.13
negative	0.15	0.81	0.04
neutral	0.18	0.17	0.75
Sentence level			
True/Estimated	positive	negative	neutral
positive	0.62	0.19	0.19
negative	0.09	0.86	0.05
neutral	0.30	0.41	0.29

Misclassification of sentences/documents are due to different reasons such as lack of background knowledge. A sample misclassified sentence is provided below.

5 puan verdim, o da janistonun güzel yüzünün hatırına.
(I gave 5 points, and that because of the lovely character of Janiston).

In this example, our system cannot distinguish "5 points" (out of 10) as a low grade for a movie and therefore misclassifies this negative sentence as positive because of the positive phrase in it.

6.4 Discussion and Comparison

As seen in Tables 8 and 9, the obtained accuracies in different cases range from 60% to 79%. Considering the results, we came up with the following conclusions:

- Document level sentiment analysis is more successful compared to sentence level, as expected. The intuition is that correctly classified sentences in a document compensate for misclassified sentences.
- The most effective group of features at binary sentence level task are $f_5 - f_6$ (number of positive and negative words in PWS). As PWS contains only positive and negative words, these features are not very effective in ternary classification.
- The most effective features at document level in isolation are $f_{17} - f_{20}$ (polarity of the first and last sentences). This observation is in agreement with the assumption that the first and last sentences in a document are the best estimators of the document polarity. This was also cited in literature by a few researchers such as Meena and Prabhakar (2007) and Gezici et al. (2012). In fact, the difference in classification accuracy between using only the polarity of the first and last sentence, and using all features (in document level) is less than one percentage point.
- The least effective feature set in isolation is $f_7 - f_8$ (emoticons) for both sentence and document level analyses.
- In almost all settings, each added feature subset improves the accuracy over the existing features. For example, adding $f_{17} - f_{20}$ to feature group $f_1 - f_{16}$, increases the accuracy by one percentage point.
- Generally, our system is more successful in classifying positive sentences and documents compared to negative or neutral ones.
- Our approach improves upon the simple baseline of using average word polarities (features $f_1 - f_2$) in the review by about four percentage points.

We could not apply other methods in the literature on our dataset because none of the previous works have released their detailed approach or polarity lexicons. Moreover, related research report only binary classification results which neglects neutral reviews, while we consider both binary and ternary classifications.

Similar works to ours are (Vural *et al.* 2012) and (Eroğul 2012), which have reported 76% and 85% accuracy in classifying Turkish movie reviews as positive and negative. The comparable accuracy (binary document classification) in our work is

79%; however these accuracies may not be directly comparable as the details of how they used the dataset are unknown. Moreover, previous work focus on document level sentiment analysis, while we consider aspect and sentence levels as well.

7 Conclusions and Future Work

Despite the fast growth of sentiment analysis in English, most other languages suffer from a shortage of research in this area. We designed and implemented a comprehensive sentiment analysis system that uses several resources and achieves a good accuracy by handling many linguistic issues for Turkish. Although we were unable to address all sub-problems in Turkish sentiment analysis, we comprehensively defined the problem to clarify those issues that need more attention in the future. Our system works in different levels, namely aspect, sentence, and document levels.

We will extend our system in the future by (1) investigating phrase level sentiment analysis more deeply, (2) effectively addressing more language issues such as conditional sentences, and (3) extending the current system for other domains and data types, such as tweets.

References

Akın, A. A. and Akın, M. D. 2006. Zemberek, an open source NLP framework for Turkic Languages *Structure 10.*

Bilgin, O., Çetinoğlu, Ö. and Oflazer, K. 2014. Building a wordnet for Turkish. In *Romanian Journal of Information Science and Technology* (7) 1–2 pp. 163–172.

Boynukalın, Z. 2012. Emotion analysis of Turkish texts by using machine learning methods. *MSc thesis, Middle East Technical University, Turkey*

Cambria, E., Olsher D. and Rajagopal D. 2014. SenticNet 3: a common and common-sense knowledge base for cognition- driven sentiment analysis. *Twenty-eighth AAAI conference on Artificial Intelligence*, pp. 1515–1521.

Dehkharghani, R., Saygin, Y., Yanikoglu, B. and Oflazer, K. 2014. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. Technical Report, Faculty of natural science and engineering, Sabanci University.

Demiröz, G., Yanikoglu, B. Tapucu, D. and Saygin, Y. 2014. Learning domain-specific polarity lexicons. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* (7) pp. 674–679.

Eroğul, Umut. 2012. Sentiment analysis in Turkish. *MSc thesis, Middle East Technical University, Turkey.*

Eryiğit, Gülşen. 2014. ITU Turkish NLP Web Service. In *proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden* pp. 1–4.

Esuli, A. and Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* 6: pp. 417–422.

Gezici, G., Yanikoglu, B., Tapucu, D. and Saygın, Y. 2012. New Features for Sentiment Analysis: Do Sentences Matter?. *SDAD 2012, the first International Workshop on Sentiment Discovery from Affective Data*

Hatzivassiloglou, V. and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.*

Hosmer Jr, D. and Lemeshow, S. 2014. Applied logistic regression. *John Wiley and Sons, New Jersey, USA.*

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*

Kaya, M., Fidan, G. and Toroslu, I. H. 2012. Sentiment Analysis of Turkish Political News. In*proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology* (1) pp. 174–180

Liu B. 2012. Sentiment Analysis and Opinion Mining. *Morgan and Claypool Publishers, USA.*

Meena, A. and Prabhakar, TV. 2014. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis.*Springer Berlin Heidelberg*, pp. 573–580.

Seslisözlük Group. 2014. Sesli sözluk, bilingual English-Turkish-English dictionary. http://www.seslisozluk.net *Accessed: September 2014.*

Thelwall, M. and Buckley K. Paltoglou, G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*

Vural, A. G., Cambazoğlu, B. B., Şenkul, P. and Tokgöz, Z., Ö. 2012. A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish. *ISCIS* pp. 437–445.

Yıldırım, E., Çetin, F. S., Eryiğit, G. and Temel, T. Ö. 2015. The Impact of NLP on Turkish Sentiment Analysis. *Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi* (8)(8).