Natural Language Engineering **23** (4): 535–559. © Cambridge University Press 2016
doi:10.1017/S1351324916000309

Sentiment analysis in Turkish at different granularity levels

RAHIM DEHKHARGHANI[1], BERRIN YANIKOGLU[2], YUCEL SAYGIN[2] and KEMAL OFLAZER[3]

[1]*University of Bonab, Bonab, Iran*
e-mail: rdehkharghani@bonabu.ac.ir
[2]*Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey*
e-mails: berrin@sabanciuniv.edu, ysaygin@sabanciuniv.edu
[3]*Carnegie Mellon University – Qatar, Doha, Qatar*
e-mail: ko@cs.cmu.edu

(*Received 28 August 2015; revised 19 November 2016; accepted 13 September 2016; first published online 21 October 2016*)

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. We cover different levels of sentiment analysis such as aspect, sentence, and document levels as well as some linguistic issues such as conjunction and intensification in Turkish sentiment analysis. Our system is evaluated on Turkish movie reviews and the obtained accuracies range from sixty per cent to seventy-nine per cent in ternary and binary classification tasks at different levels of analysis.

1 Introduction

Sentiment analysis or 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, and political platforms, providing an opportunity for companies and organizations to gather feedback from a large group 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* has a positive polarity for *room size* in the hotel domain but a negative polarity for *battery size* in the camera domain. Similar examples may be given for the same concept being neutral versus polar across 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.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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 the Turkish language, 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 document level. More research is required in order to address different levels of granularity (e.g., aspect level) and support ternary classification handling neutral/objective opinions.

In earlier work, we have built polarity resources for Turkish such as SentiTurkNet (Dehkharghani *et al.* 2016), and a polar word set consisting of 2,000 words (Dehkharghani *et al.* 2014). In this work, we propose a sentiment analysis system for Turkish and apply it on Turkish movie reviews.[1] The system uses previously built polarity resources for extracting various features which are used in a supervised learning approach. While statistical approaches may work better when there is sufficient amount of data in the given domain, our framework is applicable to all languages for which there is a domain-independent lexicon, with a small amount of effort to collect a seed word list of domain-specific words. Our contributions can be summarized as follows:

- We provide an overview of issues for sentiment analysis in Turkish.
- We propose and evaluate a sentiment analysis system for Turkish covering linguistic issues and different levels of analysis granularity. This system exploits polarity lexicons such as SentiTurkNet and additional natural language processing (NLP) techniques such as dependency parsing that have not yet been employed for Turkish sentiment analysis in the literature.

The remainder of this paper is structured as follows: Section 2 explains the characteristics of Turkish and its challenges in sentiment analysis. Sections 3 and 4, respectively explain various granularity levels and linguistic issues in sentiment analysis. Section 5 reports previous work on sentiment analysis, mostly in Turkish and partially in English languages. Section 6 introduces our proposed system for Turkish. The proposed system is evaluated in Section 7 which is followed by conclusions and future work in Section 8.

2 Turkish and its challenges in sentiment analysis

Turkish is a member of the Turkic family of Altaic languages. Particular characteristics of Turkish make NLP and sentiment analysis tasks difficult for this language. Morphologically, Turkish is an agglutinative language with morphemes attaching to a root word as 'beads-on-a-string'. Words are formed by very productive affixations of multiple suffixes to root words, from a lexicon of about 30 K root words (not

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

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

counting proper names.) Nouns do not have classes nor are there any markings of grammatical gender in morphology and syntax. When used in the context of a sentence, Turkish words can take many inflectional and derivational suffixes. It is quite common to construct words which correspond to almost a sentence in English: For example, the equivalent of the Turkish word *sağlamlaştırabileceksek* in English can be expressed with the fragment *if we will be able to make [it] become strong (fortify it)* (Oflazer and Bozşahin 1994).

For Turkish, the morphological structure of a word is also necessary for sentiment analysis in addition to the root word, as suffixes may change the polarity of a word. For instance, the word *iştahsız* (having no appetite), is negative (due to suffix *+sız*), while its antonym, *iştahlı*, is positive (due to suffix *+lı*). Note that the root word itself, *iştah* (appetite), is mostly positive.

In terms of NLP tools for Turkish, there are morphological analyzers (Oflazer 1994; Akın and Akın 2007), and parsers (Eryiğit 2014). Morphological analyzers perform a word-level analysis and provide the root word. We review the parser in Section 6.1 as it is relevant to our work here.

3 Granularity levels in sentiment analysis

The most common level of sentiment analysis is 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) in a document is positive but another aspect (e.g., director) in the same document is negative, the sentiment analyser may classify this document as neutral while in fact it has mixed sentiment. A 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 in Section 1).

Syntactic structure level: At this level, we try to estimate sentiments of syntactically related constituents of a sentence which is later used in aspect-level sentiment. For example, the sentence below has two components with two different sentiments (one is shown in italic):

Ben beğendim, *ama herkes beğenmedi.*
(I liked it, *but not everyone did.*)

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 differ from the overall polarity of the sentence. For example, the sentence below has two parts about two separate aspects, one with positive and the other with negative sentiment (one is shown in italic).

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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

Sentence level: Sentiment analysis of sentences may follow word and syntactic structure level analyses. If a sentence has a mixed polarity (both positive and negative due to multiple aspects or syntactic structures), one can assign an overall polarity based on relative sentiment strengths of the components/phrases inside the sentence.

Document level: This is the coarsest level which 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 works (Meena and Prabhakar 2007; Gezici *et al.* 2012) show 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 NLP issues in sentiment analysis

In order to make more precise sentiment classifications, a sentiment analysis system should handle various linguistic markers such as negations, intensifications, and conditional constructions. Most of these markers are language-specific and their extraction requires language-specific tools (e.g., morphological analyzers and parsers), while some others such as 'background knowledge' are considered language-independent.

Below we group the issues that we rely on for sentiment analysis, into two subsets: 'linguistic' and 'other' issues. Here, we present only the challenges, while proposed solutions are presented in Section 6.3.

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 sorry.*)

where there is a negative to positive change in the sentiment as *pişman* (sorry) 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)*

In this sentence, the polarity first switches to negative with *sevmedim* (I did not like) and back to positive within *çıkmadı* (no one came out).

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Negation has been investigated in the literature mostly for English. There are two challenging issues: (1) specifying the scope of negation, that is, all words are affected by the negation maker and (2) updating the polarity of words or phrases in the scope of negation. In order to handle the first issue, Jia, Yu and Meng (2009) propose several patterns and rules to find the scope of a negator. For the second issue, Wilson, Wiebe and Hoffmann (2005) shift the polarity score by decreasing the positivity or negativity scores of the negated part.

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*). This issue as well as its effect on negated clauses are discussed by Wilson, Wiebe and Hoffman (2005). In this work, the authors propose to increase or decrease the polarity score of a polar word proceeded after an intensifier.

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.)

One of the most comprehensive research works on sentiment analysis of conditional sentences has been accomplished by Narayanan, Liu and Choudhary (2009), in which, conditional sentences are divided into four types based on the verb tense. Then, features such as opinion words, their position in the sentence, their POS tags, and conditional connectives are extracted from each sentence.

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 message or sentiment. 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 a negative polarity.

İnsan bu filmi sevmeyebilir *mi*?
(*Can anyone 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 (fifty-seven per cent) even in English (Liu 2012). In Turkish, a sarcastic statement such as *harika bir film olmuş!* (it was a great movie!) at the end of a pure negative review, can only be detected by the disagreement with it and the whole of the (negative) review and slightly hinted by the exclamation mark.

Davidov, Tsur and Rappopor (2010) propose a semi-supervised method for recognition of sarcastic sentences in Twitter and Amazon reviews. The proposed method includes two phases: pattern acquisition to identify sarcastic patterns to

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

enable feature extraction, and classification stage which classifies each sentence into sarcastic or non-sarcastic classes.

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 automatically extracted by using the polarity of its parts. 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 constituents *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 require attention by sentiment analysis systems across different languages.

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 conjoined by *ve* (and). This observation was made and used to estimate word-level polarities by Hatzivassiloglou and MCkeown (1997).

Background knowledge: Sentiment analysis systems require background (common-sense) knowledge for classifying special kinds of sentences such as 'one of those rare films that makes me feel like I'm there'. 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 NLP to extract such information. While there are common-sense knowledge resources such as Cyc (Lenat and Guha 1989) and ConceptNet (Havasi, Speer and Alonso 2007) for English, there are only few works to build (Cambria *et al.* 2012) or incorporate common-sense knowledge in different recognition/understanding tasks, including sentiment analysis or language understanding (Sureka *et al.* 2009; Wang *et al.* 2013).

5 Related work

English is the richest language in terms of the amount of research for sentiment analysis. Pang, Lee and Vaithyanthan (2002) investigated document level sentiment analysis by using machine learning approaches. In sentence level sentiment analysis, Meena and Prabhakar (2007) addressed the effect of conjunctions, and semantic relations between sentences. In aspect level sentiment analysis, Ding, Liu and Yu (2008) estimate the polarity of aspects by analyzing the polarity of neighbor words for each aspect in a window. Finally, in phrase level sentiment analysis, we can cite the work of Wilson, Wiebe and Hoffmann (2005), and Wilson, Wiebe and Hoffmann

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

(2009) which first classify an expression as subjective or objective and then estimate its polarity in the case of subjectivity.

There are relatively few works in sentiment analysis of Turkish text. Yıldırım *et al.* (2015) perform sentiment analysis on Turkish tweets in the telecommunications domain. They applied a multi-class ternary (positive, negative, and neutral) classification with support vector machines on tweets using features such as inverse document frequency, unigrams, and adjectives. They also employed NLP techniques such as normalization, stemming, and negation handling. The best accuracy in classifying tweets is reported as seventy-nine per cent.

Türkmenoğlu and Tantuğ (2014) compared lexicon-based and machine learning-based sentiment analysis methods on Turkish social media. In the lexicon-based approach, the authors used the translation of an English opinion lexicon. The authors also used a machine learning approach with different feature sets including word roots, n-grams, and bag of words. Both approaches in this paper were applied on binary (positive/negative) classification of Turkish movie reviews. The classification accuracy of their method in the binary classification (positive and negative) is 75.2 per cent for Twitter dataset and 79.0 per cent for movie dataset.

Kaya, Fidan and Toroslu (2012) investigated sentiment analysis of Turkish political news in online media. They use four different classifiers – Naive Bayes, maximum entropy, SVM, and character-based n-gram language models – with a variety of text features: frequency of word unigrams, bigrams, root words, adjectives, and effective (polar) words. They conclude that the maximum entropy and n-gram language models are more effective than the SVM and Naive Bayes classifiers in classifying Turkish political news as positive and negative. They report an accuracy of seventy-six per cent in binary classification of political news. Later, Kaya, Fidan and Toroslu (2013) used transfer learning applied from unlabeled Twitter data to labeled political columns. A variation of self-taught learning has been proposed for classification in this work, where the goal is also to determine if a Turkish political document is positive or negative, neglecting the neutral documents. In this work, a simple baseline classifier technique gives fifty-nine per cent accuracy and the proposed method obtains up to twenty-six per cent increase in the accuracy of classification with the inclusion of Twitter data into classification of Turkish political columns.

Vural *et al.* (2013) present a system for unsupervised sentiment analysis in Turkish text documents by customizing SentiStrength (Thelwall, Buckley and Paltoglou 2012) by translating its polarity lexicon to Turkish. SentiStrength is a sentiment analysis tool for English which assigns a positive and a negative score to a segment of text. The authors report seventy-six per cent accuracy in classifying Turkish movie reviews as positive and negative, using this method.

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. Boynukalın reports an accuracy of seventy-eight per cent in classifying documents into those four emotions.

Eroğul (2012) explores the use of surface linguistic features such as part-of-speech (POS) tags, word unigrams and bigrams, and negation markers in sentiment analysis.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

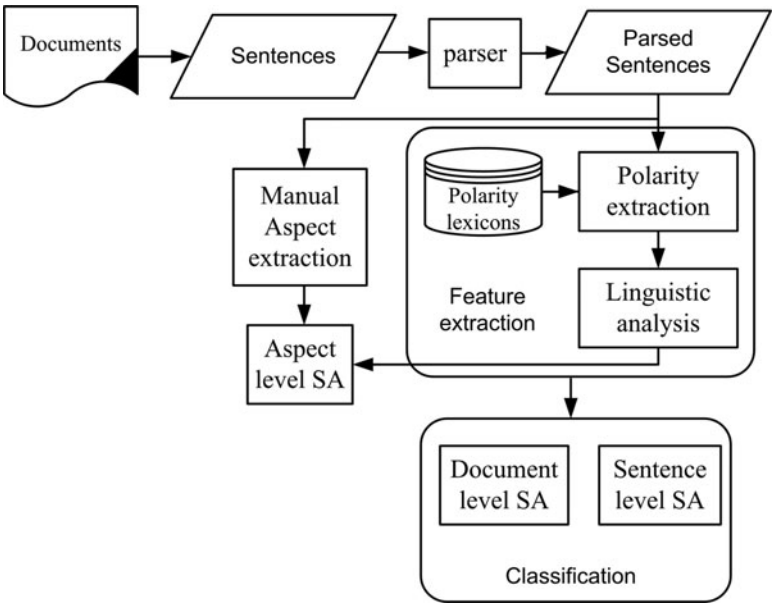


Fig. 1. The proposed system as a flowchart.

This work relies on a morphological analyser for Turkish, called *Zemberek* (Akın and Akın 2007), and reports an accuracy of eighty-five per cent on classifying Turkish movie reviews (at 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.

The existing works on sentiment analysis of the Turkish language are limited in that the suggested systems do not consider all granularity levels or all linguistic issues. Furthermore, the results are evaluated on proprietary or small datasets, which makes it difficult to compare various approaches with each other. In this article, 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.

6 Our methodology

In this section, we first present an overview of our system and then elaborate on each component of the system. 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 a morphological analysis for each word. This structure is used in aspect-level polarity classification (see Section 6.2.3).

We assign polarity scores to word n-grams (unigram, bigram, and trigram) in the component 'polarity extraction' by using the following polarity lexicons: SentiTurkNet, Polar Word Set, and the translation of SenticNet (see Sections 6.1 and 6.2.4).

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 1. *Parse tree generated by using the parser for the sentence Bence hoş vakit geçirmek için seyredilebilir.* (*It could be seen for an enjoyable time*)

Position	Word	Root	Major POS	Minor POS	Morph. Features	Dep. Head
1	Bence	ben	pron	pers	A1sg.pnon.equ	7
2	hoş	hoş	adj	adj	–	3
3	vakit	vakit	noun	noun	A3sg.pnon.nom	4
4	–	geçir	verb	verb	Pos	5
5	geçirmek	–	noun	Inf1	A3sg.pnon.nom	6
6	için	için	postp	pcnom	–	7
7	seyredilebilir	seyredil	verb	Able	Pos.aor.a3sg	0

We handle various linguistic issues in the 'Linguistic analysis' component. Then, features are extracted from each sentence and each document to train a sentence-level and a document-level classifier. The trained classifiers are used to perform a sentence-level and a document-level polarity classification in components 'Sentence level SA' and 'Document level SA'.

The proposed approach either benefits from the existing techniques in the literature to deal with sentiment analysis issues in Turkish, or proposes new solutions for them. In order to apply this framework on other languages, some language-specific changes need to be made. First, the proposed framework relies on polarity lexicons, which makes those lexicons crucial for using this framework. Furthermore, some linguistic issues such as negation handling are language-specific and need to be handled by different methods in each language.

6.1 NLP tools and polarity resources

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

ITU Turkish Parser (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 could be seen for an enjoyable time) is illustrated in Table 1. This tool provides tokenization, normalization, spelling correction, morphological analysis, dependency parsing, and few other features. After manually evaluating the output of this tool for a small subset of sentences, we conclude that it is quite successful in tokenization, normalization, and spelling correction, while its accuracy is worse in parsing. Nonetheless, we opted to use it as there is no other tool that provides comparable functionality. Our approach in this work is to use multiple resources and features to obtain a robust sentiment analysis system in the presence of such errors.

Polar Word Set (PWS): We have semi-automatically generated a list of polar Turkish terms including 1,000 positive and 1,000 negative terms using the method

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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. This lexicon is used in the sentence and document level sentiment analyses.

*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 (Dehkharghani *et al.* 2014). This resource consists of about 15,000 synsets and 1.47 terms per synset in average. This lexicon is used in the aspect, sentence, and document level sentiment analyses.

*SenticNet (*SN*):* This resource assigns various values to each term as its *pleasantness, attention, sensitivity, aptitude* scores, as well as an *overall-polarity*. Each of these scores has a value between -1 and $+1$ as the most negative and the most positive polarities, respectively. We translated this resource to Turkish using a bidirectional dictionary named *SesliSözlük* (Seslisözlük Group 2014) and used only the *overall polarity* of each term (or phrase) as its sentiment polarity. This lexicon contains about 14,000 entries (words and phrases). This lexicon is used in sentence- and document-level sentiment analyses. Note that translating English resources to other languages is problematic as the sentiment polarity cannot always be transferred using translation from one language to another. In particular, language-dependent or cultural-dependent words such as *Tanrı* (God), may have different polarities in different cultures (Dehkharghani *et al.* 2014). The existence of such words can affect the system negatively; however, these problem cases are relatively rare within a review.

Domain-indicative keywords: This resource consists of ten keywords and key phrases which are used in feature f_{11} as a small list of seed words (see Section 6.4). Appearance of these keywords in a review indicates a positive polarity for it in the movie domain. This lexicon is used in the sentence- and document-level sentiment analyses. This list is given below:

izleyin/izle (watch it), *iyi seyirler* (happy viewing),*kaçırma/kaçırmayın* (do not miss it), *izlenmeli/izlemeli/izlemek gerek/izlemek lazım* (should be watched), *izlenebilir* (could be watched).

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. The word *güzel* (good/beautiful), for example, is labeled 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.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

6.2 Sentiment analysis for Turkish at various granularity levels

Our system is designed to address different levels of sentiment analysis: word, aspect, sentence, and the overall document, as explained below.

6.2.1 Word level

We extract the polarity of a given word using the polarity lexicons described in Section 6.1 (*PWS*, *SN*, and *STN*). Polarities of constituent words are used in all subsequent steps, including feature extraction of the sentence- and document-level analyses.

Since *STN* is derived from the Turkish WordNet, different senses of a word may be associated with different polarity scores. The proper approach would have been to perform Word-Sense Disambiguation to find the correct sense of the word considering the context. Word-Sense Disambiguation 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 *POS* tags. 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 6.3. The modified polarity scores/labels are used as word polarities in all subsequent steps (aspect, sentence, and document-level classifications).

6.2.2 Exploiting dependency chains

We use the dependency parse tree constructed by the parser described in Section 6.1 to extract dependency chains. These dependency chains capture syntactic relations between words in a sentence, that are used to capture aspect level sentiment. We generate all dependency chains with two or more words that are structurally related together in the dependency parse tree (according to the last column in Table 1). There are no constraints that limits this construction (e.g., POS tag limits). For example, in the parsed sentence of Table 1, the term *hoş* (enjoyable) is related to (dependent on) the term *vakit* (time), *vakit* is related to *geçirmek* (spending), *geçirmek* is related to *için* (in order to), and finally *için* is related to *seyredilebilir* (can be viewed); so the dependency chain *hoş vakit geçirmek için seyredilebilir* (It can be viewed in order to spend an enjoyable time) is extracted from the sentence. However, all subsets of this dependency chain, such as *hoş vakit geçirmek için*, are also extracted. Note that the number k in the last column of Figure 1 for the term t_i indicates that t_i is related to t_k.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

6.2.3 Aspect level

As mentioned in Section 3, aspects refer to different perspectives relating to the item being reviewed. For instance, 'scenario' and 'acting' form two separate aspects of a movie. Each aspect may have a different polarity than the overall polarity. Text that refer to a particular aspect are identified by detecting keywords related to that aspect. For instance, the words 'plot', "scenario', 'story', and 'ending' all refer to the scenario aspect. We manually compiled a list of aspects (A) in the 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. We compute average polarity (positivity and negativity) of all chains involving an aspect a_j in a sentence by means of two terms $P(a_j)$ and $N(a_j)$ that indicate the average positivity and negativity scores of aspect a_j based on STN lexicon. These polarities are computed by using equations (1) and (2).

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

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

where NG is the set of all chains generated by the dependency parse tree; and n_k is a chain 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 somewhere in the chain of dependants and heads constituting n_k, as extracted from STN. Note that an aspect a_j is a member of a chain n_k ($a_j \in n_k$), if a_j appears in the dependency chain n_k.

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 below:

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

An example sentence is given below:

Senaryo güzel, *ama efektler çok kötü.*
(The scenario is good, *but the special effects are very bad.*)

The parse tree of the above-mentioned sentence is provided in Table 2. In this sentence, the two parts *senaryo güzel* (The scenario is good) and *efektler kötü* (special effects are bad) are extracted from the dependency tree, ignoring other words that do not affect the sentiment. Note that although the word *çok* (very) is not included in the chain *efektler kötü*, intensification handling doubles the polarity of adjective (*kötü*), which in turn doubles the polarity of the aspect (*efekt*) modified by the adjective. As a side note, a misleading chain *senaryo güzel kötü* (the scenario is good bad!) is also generated, but it does not change the positive polarity of the aspect *Senaryo* (Scenario) because we use average polarity of all chains including the aspect.

We believe that taking into account only neighboring words to extract the polarity of an aspect is rather simplistic. For example, consider the above-mentioned sentence:

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 2. *Parse tree generated by using the parser for the sentence Senaryo güzel, ama efektler çok kötü (The scenario is good, but the special effects are very bad)*

Position	Word	Root	Major POS	Minor POS	Morph. Features	Dep. Head
1	Senaryo	senaryo	Noun	Noun	A3sg.Pnon.Nom	2
2	güzel	güzel	Adj	Adj	–	6
3	ama	ama	Conj	Conj	–	2
4	efektler	efekt	Noun	Noun	A3pl.Pnon.Nom	6
5	çok	çok	Adverb	Adverb	–	6
6	kötü	kötü	Adj	Adj	–	0

senaryo güzel, ama efektler çok kötü (The scenario is good, but the special effects are very bad/weak). Here, taking into account five neighbor words (two words from each side) would result in associating the word *güzel* (good/beautiful) with the second aspect (*efekt*), while the proposed method extracts the chain *efektler kötü* (special effects are bad/weak), which can estimate the negative polarity of the aspect *efekt(ler)* (special effect(s)).

Note that in the aspect level analysis, we also normalize the words and apply NLP techniques such as negation and intensification handling, as also done in the sentence level. For example, in the above mentioned sentence, the aspect word *efektler* (special effects) is normalized as *efekt* (special effect).

6.2.4 Sentence and document levels

We start sentence-level sentiment analysis by automatically segmenting each document to its sentences by using punctuation, capitalization, and emoticons. Then, we extract sixteen features, explained in Section 6.4, from each sentence to be used in the classification task. The classifier is trained with feature vectors extracted from 2,700 sentences labeled as positive, negative, or neutral in the Turkish movie reviews and evaluated by five-fold cross validation.

We address document level sentiment analysis similar to the sentence level analysis, using twenty features explained in Section 6.4. The classifier is trained by 1,000 feature vectors which have been extracted from 1,000 documents labeled as positive, negative, or neutral in the Turkish movie reviews. In this level, we benefit from polarity of the first and last sentences to highlight the effect of the first and last sentences in the document. The evaluation method for this classifier is again five-fold cross validation.

6.3 Handling NLP 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, chains, and sentences to change their polarity if applicable. The initial polarity scores and labels for words

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

are obtained from three polarity lexicons as explained in Section 6.1. Although issues discussed in this section have been addressed in some of the previous works, our work addresses all major issues in one system. Also, addressing linguistic issues (e.g., negation detection and handling) requires different techniques in different languages, as provided in this work for Turkish.

6.3.1 Linguistic issues

In this section, linguistic issues which are mostly language-dependent are discussed as it applies to Turkish. Due to the nature of natural languages, the issues mentioned below should be covered in different ways, for different languages. For example, in Turkish, negation marker attaches to the verb but in English, it comes before the verb.

Negation: We covered different kinds of negation in Turkish which are listed below:

- The predication negation marker *değil* (is/am/are not) switches the sentiment of the preceding 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 a positive sentiment but the 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 parser.
- Morphemes *+lu* and *+suz* (and their all morphemes) derive adjectives from nouns 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.

Intensification: We compiled a set of intensifiers in Turkish listed below. For strengthening intensifiers, we double the sentiment value and for weakening intensifiers we halve it (There should be an upper limit for the polarity score, e.g., 0.95).

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

Conditional sentences: We cover this only by adding a boolean feature to the classification features 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.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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.

6.3.2 Other issues

In this work, we covered only the conjunction issue from Section 4.2:

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) in the movie domain as negative because it already knows that *güzel* (beautiful/good) is positive.

6.4 Features for sentence and document classification

We explain the sixteen and twenty features used in sentiment classification of sentences and documents in this section. Features $f_1 - f_{16}$ in two tables are similar (the input is either a sentence or a document), but $f_{17} - f_{20}$ are additional features used only in document level. Below we explain these features in some detail. The term 'review' used for feature explanations refers to a sentence or document, in the sentence- and document-level sentiment analyses, respectively. We believe that some features in this work are novel – such as the sentence type features ($f_{13} - f_{16}$). Furthermore, features based on SentiTurkNet – the first Turkish polarity lexicon – are evaluated for the first time.

Word polarities ($f_1 - f_4$): The first four features capture the average polarity of terms in a review, computed using two separate resources which 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.

Polar words ($f_5 - f_6$): These features indicate the number of positive and negative terms in each review, as computed according to the *PWS*.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 3. *Features used in sentiment analysis of a sentence, S. SN, PWS, and STN, respectively stand for SenticNet, PolarWordSet, and SentiTurkNet*

f_{s1}: *average positive score of words in S using STN*
f_{s2}: *average negative score of words in S using STN*
f_{s3}: *average score of positive words in S using SN*
f_{s4}: *average score of negative words in S using SN*
f_{s5}: *number of positive words in S using PWS*
f_{s6}: *number of negative words in S using PWS*
f_{s7}: *occurrence of positive emoticons in S*
f_{s8}: *occurrence of negative emoticons in S*
f_9: *number of adjectives and adverbs in S*
f_{s10}: *number of capital initial words in S*
f_{s11}: *number of domain-specific indicative words in S*
f_{s12}: *length of sentence (number of tokens in S)*
f_{s13}: *is S a conditional sentence?*
f_{s14}: *is S an interrogative sentence?*
f_{s15}: *is S a negated sentence?*
f_{s16}: *is S an exclamative sentence?*

Table 4. *Features used in sentiment analysis of a document, D. SN, PWS, and STN, respectively stand for SenticNet, PolarWordSet, and SentiTurkNet*

f_{d1}: *average positive score of words in D using STN*
f_{d2}: *average negative score of words in D using STN*
f_{d3}: *average score of positive words in D using SN*
f_{d4}: *average score of negative words in D using SN*
f_{d5}: *number of positive words in D using PWS*
f_{d6}: *number of negative words in D using PWS*
f_{d7}: *occurrence of pos. emoticons in D*
f_{d8}: *occurrence of neg. emoticons in D*
f_{d9}: *number of adjectives and adverbs in D*
f_{d10}: *number of capital initial words in D*
f_{d11}: *number of domain-specific indicative words in D*
f_{d12}: *length of document (number of tokens in D)*
f_{d13}: *does D contain a conditional sentence?*
f_{d14}: *does D contain an interrogative sentence?*
f_{d15}: *does D contain a negated sentence?*
f_{d16}: *does D contain an exclamative sentence?*
f_{d17}: *avg. positive score of words in first sentence of D*
f_{d18}: *avg. negative score of words in first sentence of D*
f_{d19}: *avg. positive score of words in last sentence of D*
f_{d20}: *avg. negative score of words in last sentence of D*

Emoticons ($f_7 - f_8$): These features indicate the number of positive and negative emoticons in the review. Appearance of positive (or negative) emoticons indicate a positive (or negative) polarity for the review including the emoticon.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 5. *Conditional probability of the document polarity given the polarity of the first or last sentence*

Document	P(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	P(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

Linguistic issues ($f_9 - f_{12}$): These features are defined based on our three assumptions: (1) the higher the number of adjectives and adverbs in a review, the greater the chances of its subjectivity; (2) the higher the number of initial capital words in a review, the greater the chances of neutrality in 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. For feature f_{11}, we gathered a short list of keywords and key phrases by simply observing the data, indicating some positive sentiment statements in Turkish movie reviews which was listed in Section 6.1.

Sentence type ($f_{13} - f_{16}$): These features capture the interrogative, conditional, negated, or exclamative form of a sentence. In document level, the existence of such sentences in a document is used. These features can be extracted from the output of the parser. These features affect the sentiment of a sentence or document in some way. For example, interrogative and exclamative sentences may have a higher potential to be subjective.

First and last sentences ($f_{17} - f_{20}$): Polarities of the first and last sentences in a 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 analyzed the relationship between the document polarity and the polarity of its first and last sentences. Table 5 shows the conditional probabilities of the document polarity given the sentence polarity. For instance, seventy-six per cent of documents with positive sentiment have a positive first sentence. As shown in previous work, these numbers also indicate that the first sentence polarity is especially indicative of document polarity.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

6.5 Classifier training

For sentence classification, sixteen features in Section 6.4 are used with a logistic regression classifier (Hosmer and Lemeshow 2012). This classifier is used in all experiments. We wanted to have the outputs correspond to posterior class probabilities, which the logistic regression is trained to provide. The evaluation is done using five-fold cross validation over training data of 2,700 sentences. Both binary and ternary classifiers are trained separately at this level.

For document classification, we use the twenty features in Section 6.4. The classifier and evaluation methods are the same as in the sentence-level analysis, using logistic regression as the classifier and five-fold cross validation for evaluation.

7 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 five-fold cross-validation on training data.

7.1 Dataset

We used a subset of Turkish movie reviews as dataset and manually labeled 1,000 randomly chosen documents from the dataset as positive, negative, or neutral. We also labeled 2,700 sentences appearing in these documents as positive, negative, or neutral. Finally, we also manually labeled all aspects in these sentences, which resulted in about 2,000 aspect mentions being labeled as positive, negative, or neutral. The label annotation process is done by three people and the degree of agreement among all three labelers was eighty-six per cent for document level, eighty-four per cent for sentence level, and eighty-two per cent for aspect level analysis. We did not use the already assigned rating scores to each movie review because we need labels for the aspect, sentence, and document levels and the existing rating scores of movie reviews are available only at document level. We preferred manual labeling which is also more accurate than the rating scores. Each aspect, in the aspect level sentiment analysis, is labeled as positive, negative, or neutral in a sentence, if it conveys a positive, negative, or neutral polarity to the reader.

The distribution of [positive, neutral, and negative] sentences and documents are close: [fifty per cent, thirty per cent, twenty per cent] and [fifty-two per cent, twenty-nine per cent, ninteen per cent], respectively.[2] We did not include the label *mixed* in our labeling; instead we chose the dominant sentiment in a mixed review and labeled it accordingly.

Note that the number of documents in the Turkish movie reviews is about 60,000 and previous works have all used small random subsets that are not public. Therefore, we could not evaluate our system with those datasets. However, in order to give a comparison opportunity for future works, we have made our dataset public.

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

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

7.2 Dealing with unbalanced data

As mentioned above, our dataset is unbalanced in favor of positive reviews, which causes biased results for the positive samples (sentences and documents) during the classification. To avoid this problem, we balanced the dataset by re-sampling under-represented classes. The re-sampling approach to unbalanced data has been used in many other works (Estabrooks, Jo and Japkowicz 2004; Minaei-Bidgoli, Topchy and Punch 2004). This technique increased per-class classification accuracies (by five per cent to twenty-five per cent in various cases) for the negative and neutral classes, while the per-class accuracy of the dominant (positive) class decreased by eleven per cent and four per cent in the document and sentence levels. The overall accuracy increased by one per cent points.

7.3 Results

The accuracies obtained from binary and ternary classifications on the sentence and document levels are presented in Tables 6 and 7. Using all the features, we obtained 73.42 per cent and 79.06 per cent accuracies in binary sentence and document classification problems, respectively. For ternary classification, results are 60.33 per cent and 73.01 per cent for the sentence and document levels. As expected, higher accuracies are achieved at document level (due to larger context) and binary classification problems (simpler problem). The standard deviations obtained in five-fold cross-validation experiments are indicated alongside the mean accuracies and indicate low variance with respect to different train and test splits.

The last two rows in Tables 5 and 6 are added for isolated evaluation of two linguistic issues: negation and intensification. Negation handling increases the classification accuracies by 1–2 percentage points. Intensification handling increase the accuracy only by 0.5–1 percentage point, as this issue is rare in the reviews. The effect of some issues such as emoticons and sentence types are also shown as features in the above mentioned tables.

We also performed an aspect-based sentiment analysis by using equations (1) and (2) and achieved seventy per cent and seventy-nine per cent accuracies in ternary and binary classifications, respectively. We do not investigate the aspect-level sentiment analysis in depth, unlike the sentence and document levels. We rather use a rule-based method which uses two proposed equations to compute the average positivity and average negativity scores of neighbor words of an aspect in the dependency tree, to estimate the polarity of the aspect in the sentence.

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 four percentage points, at document level (75.04 per cent versus 79.06 per cent and 69.30 per cent versus 73.01 per cent). In the sentence level, the same improvement in binary and ternary classifications are fourteen per cent and one per cent.

The confusion matrices for both binary and ternary classifications are given in Table 8. Each value in this table shows the per-class accuracy (diagonal values in

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 6. *Sentence level binary and ternary classification accuracy (%) by logistic regression using five-fold cross validation. The value after ± is the standard deviation over five-fold cross validation*

Feature subset	Binary	Ternary
STN features: $f_{s1}-f_{s2}$	50.73 ± 0.88	59.33 ± 0.46
SN features: $f_{s3}-f_{s4}$	59.00 ± 0.16	58.74 ± 0.12
PWS features: $f_{s5}-f_{s6}$	63.24 ± 1.07	59.61 ± 0.94
Emoticons: $f_{s7}-f_{s8}$	51.79 ± 1.18	49.20 ± 1.66
Linguistic features: $f_{s9}-f_{s12}$	51.50 ± 1.01	59.20 ± 0.93
Sentence types: $f_{s13}-f_{s16}$	57.99 ± 0.17	59.07 ± 0.20
$f_{s1}-f_{s4}$	59.73 ± 1.08	60.00 ± 1.13
$f_{s1}-f_{s6}$	70.05 ± 0.99	60.12 ± 1.05
$f_{s1}-f_{s8}$	70.40 ± 0.31	60.08 ± 0.21
f_1-f_{12}	72.2 ± 1.16	60.14 ± 1.27
All : $f_{s1}-f_{s16}$	**73.42 ± 1.60**	**60.33 ± 1.46**
All without negation	72.00 ± 0.69	59.14 ± 0.67
All without intensification	71.93 ± 1.59	60.33 ± 1.46

Table 7. *Document level binary and ternary classification accuracy (%) by logistic regression using five-fold cross validation. The value after ± is the standard deviation over five-fold cross validation*

Feature Subset	Binary	Ternary
STN features: $f_{d1}-f_{d2}$	75.04 ± 2.06	69.30 ± 2.10
SN features: $f_{d3}-f_{d4}$	76.57 ± 1.22	70.70 ± 1.08
PWS features: $f_{d5}-f_{d6}$	75.68 ± 0.73	70.61 ± 0.99
Emoticons: $f_{d7}-f_{d8}$	51.01 ± 1.43	48.42 ± 1.10
Linguistic features: $f_{d9}-f_{d12}$	74.15 ± 0.19	69.12 ± 0.29
Sentence types: $f_{d13}-f_{d16}$	73.50 ± 0.84	69.10 ± 0.51
First and last sentences: $f_{d17}-f_{d20}$	78.02 ± 0.67	72.30 ± 0.68
f_1-f_4	77.44 ± 1.09	71.10 ± 1.48
$f_{d1}-f_{d6}$	77.50 ± 1.15	71.22 ± 1.60
$f_{d1}-f_{d8}$	78.25 ± 1.90	71.2 ± 1.58
$f_{d1}-f_{d12}$	78.42 ± 1.07	71.34 ± 1.02
$f_{d1}-f_{d16}$	78.64 ± 0.29	71.51 ± 0.40
All : $f_{d1}-f_{d20}$	**79.06 ± 0.71**	**73.01 ± 0.59**
All without negation	77.90 ± 0.90	72.33 ± 0.71
All without intensification	79.00 ± 1.65	72.99 ± 1.37

matrix), separately for positive, negative, and neutral classes in ternary classification and for positive and negative classes in binary classification.

7.4 Discussion and Comparison

In the aspect-level classification, we compared the proposed method with a baseline which considers only the neighbor words around an aspect in a window with size of five words (two words from each side). We also tested the proposed approach (in the

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Table 8. *Confusion matrices for binary and ternary classifications of sentences and documents*

Document level	Binary		Ternary		
True/Estimated	**pos.**	**neg.**	**pos.**	**neg.**	**neut.**
pos	0.86	0.14	0.67	0.20	0.13
neg.	0.27	0.73	0.15	0.81	0.04
neut.	–	–	0.18	0.17	0.75
Sentence level	Binary		Ternary		
True/Estimated	**pos.**	**neg.**	**pos.**	**neg.**	**neut.**
pos.	0.92	0.08	0.62	0.19	0.19
neg.	0.67	0.33	0.09	0.86	0.05
neut.	–	–	0.30	0.41	0.29

Table 9. *A comparison on approaches to the aspect-level sentiment analysis based on classification accuracy (%)*

Movie domain		
	Proposed method	**Baseline**
Binary	79.56	75.11
Ternary	70.21	67.06
Restaurant domain		
	Proposed method	**baseline**
Binary	73.56	72.16
Ternary	68.31	67.56

aspect level) on the restaurant dataset which is a benchmark in Semeval-2016 task 5[3] including 1,233 sentences, each of which includes at least one aspect word. The achieved accuracies are reported in Table 9. As can be seen, the proposed approach improves the classification accuracy over the baseline method in the movie domain by five per cent and three per cent percentage points, respectively in binary and ternary classifications, but this improvement decreases to two per cent and 0.5 per cent in the restaurant domain.

For the sentence and document levels, as seen in Tables 6 and 7, the obtained accuracies in different cases range from sixty per cent to seventy-nine per cent. Considering the results, we can state the following conclusions:

In the sentence- and document-level classifications, the most effective group of features at binary sentence level task is *polar words* (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 $f_{d17} - f_{d20}$ (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

[3] `alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools`

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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 sentences, and using all features (in document level) is less than one percentage point. The least effective feature set in isolation is *Emoticons* for both sentence and document level analyses, because emoticons appear in only a small subset of sentences in the movie domain.

In almost all settings, each added feature subset improves the accuracy over the existing features. For example, adding *first and last sentences* 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 the negative or neutral ones, due to the assumption that people write their positive ideas more clearly than their negative ideas. 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 datasets. Moreover, related research report only binary classification results which neglects neutral reviews, while we consider both binary and ternary classifications.

Similar approaches have been done by Vural *et al.* (2013), Türkmenoğlu and Tantuğ (2014), and Eroğul (2012), which have reported seventy-six per cent, seventy-nine per cent, and eighty-five per cent accuracy in classifying Turkish movie reviews as positive and negative. These works have not accomplished a ternary classification and have not considered other levels such as the sentence and aspect levels; therefore, we could compare our approach to them only in the document level. The comparable accuracy (binary document classification) in our work is seventy-nine per cent; however, these accuracies may not be directly comparable as the details of how they used the dataset are unknown.

7.5 Error analysis

In this section, we discuss erroneous cases in classifications and other experiments such as handling linguistic issues. Misclassification of sentences/documents are due to different reasons such as lack of background (common-sense) 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 *five points* (out of ten) as a low grade for a movie and therefore misclassifies this negative sentence as positive because of the positive component in it.

An erroneous case in the aspect level is given below:

Oyunculuk olmuş mu bilemem ama *senaryo* çok eski.
(I do not know if the acting is good but the scnario is very old)

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

In this case, the 'acting' aspect has a neutral polarity but the positive word *olmuş* (good) deceives our system to estimate its polarity as positive; while due to the interrogative and negated case of the first part of the sentence, the polarity of the 'acting' aspect is neutral.

An erroneous negated case for the negation issue 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 *güzel* is out of the negation scope. Note that the comma and question mark are missed in this sentence, and without the comma, the sentence *Film güzel değil mi* would be interpreted as a question asking whether the movie is not good.

An erroneous sample for the conditional sentences 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 component *I did not like*.

An erroneous sample for the interrogative sentences is *Titanik gibi bir film nasıl sevilmeyebilir?* (How can a movie like Titanic be disliked?), in which, the classifier is unable to understand the embedded positive sentiment in the sentence. An erroneous sample for the conjunction issue is given below:

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

In this case, although two adjectives with negative polarity seem to be conjoined by *but*, they actually belong to different parts. A comma after *basit* could clarify this sentence.

8 Conclusions and future work

Despite the fast growth of sentiment analysis in English, most of 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 relatively good accuracy by handling several 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 at 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 linguistic 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. Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.556.69&rep=rep1&type=pdf`.

Bilgin, O., Çetinoğlu, Ö., and Oflazer, K. 2004. Building a wordnet for Turkish. *Romanian Journal of Information Science and Technology* **7**(1–2): 163–72.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

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

Cambria, E., Hussain, A., Durrani, T., and Zhang, J. 2012. Towards a Chinese common and common-sense knowledge base for sentiment analysis. In *Advanced Research in Applied Artificial Intelligence*, pp. 437–46. Berlin Heidelberg: Springer.

Cambria, E., Livingstone, A., and Hussain, A. 2012. The hourglass of emotions. In *Cognitive Behavioural Systems*, pp. 144–57. Berlin Heidelberg: Springer.

Cambria, E., Olsher D., and Rajagopal D. 2014. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of 28th AAAI Conference on Artificial Intelligence*, Quebec, Canada, pp. 1515–21.

Davidov, D., Tsur, O., and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the 14th Conference on Computational Natural Language Learning, Association for Computational Linguistics,* Uppsala, Sweden, pp. 107–16.

Dehkharghani, R., Saygin, Y., Yanikoglu, B., and Oflazer, K. 2016. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation* **50**(3): 667–85.

Dehkharghani, R., Yanikoglu, B., Tapucu, D., and Saygin, Y. 2012. Adaptation and use of subjectivity lexicons for domain dependent sentiment classification. In *Proceedings of Data Mining Workshops (ICDMW) 2012 IEEE 12th International Conference on*, Bruksel, Belgium, pp. 669–73.

Demiröz, G., Yanikoglu, B. Tapucu, D., and Saygin, Y. 2014. Learning domain-specific polarity lexicons. In *Proceedings of Data Mining Workshops (ICDMW) 2012 IEEE 12th International Conference on*, Bruksel, Belgium, (7) pp. 674–9.

Ding, X., Liu, B., and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In M. Najork, A. Z. Broder, and S. Chakrabarti (eds.), *Proceedings of WSDM*, Palo Alto, California, USA: ACM, pp. 231–40.

Eroğul, U. 2012. Sentiment analysis in Turkish. *MSc Thesis*, Middle East Technical University, Turkey.

Eryiğit, G. 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.

Estabrooks, A., Jo, T., and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* **20**(1): 18–36.

Esuli, A., and Sebastiani, F. 2006. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of LREC*, Genoa, Italy, (6) pp. 417–22.

Gezici, G., Yanikoglu, B., Tapucu, D., and Saygin, Y. 2012. New features for sentiment analysis: do sentences matter? In *Proceedings of SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data*, Edinburgh, Scottland, p. 5.

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 and 8th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, Madrid, Spain, pp. 174–81.

Havasi, C., Speer, R., and Alonso, J. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Proceedings of Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 27–9.

Hosmer Jr, D., and Lemeshow, S. 2014. *Applied Logistic Regression*. New Jersey, USA: John Wiley and Sons.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, pp. 168–77.

Jia, L., Yu, C., and Meng, W. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, pp. 1827–30.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press

Kanayama, H., and Nasukawa, T. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Sydney, Australia, pp. 355–63.

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*, Macau, China, (1) pp. 174–80.

Kaya, M., Fidan, G., and Toroslu, I. H. 2013. Transfer learning using twitter data for improving sentiment classification of Turkish political news. In *Information Sciences and Systems*, pp. 139–48, Springer.

Lenat, D. B. and Guha, R. V. 1989. Building large knowledge-based systems; representation and inference in the Cyc project. Addison-Wesley Longman Publishing Co., Inc.

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

Meena, A. and Prabhakar, T. V. 2014. *Sentence Level Sentiment Analysis in the Presence of Conjuncts using Linguistic Cnalysis*. Berlin Heidelberg: Springer, pp. 573–80.

Minaei-Bidgoli, B., Topchy, A., and Punch, W. F. 2004. Ensembles of partitions via data resampling. In *Proceedings of International Conference on Information Technology: Coding and Computing*, Las Vegas, Nevada, USA, pp. 188–92.

Narayanan, R., Liu, B., and Choudhary, A. 2012. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, Suntec Singapore, vol. 1(1), pp. 180–9.

Oflazer, K. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing* **2**(9): 137–48.

Oflazer, K. and Bozşahin, H. C. 1994. Turkish natural language processing initiative. In *Proceedings of the 3rd Turkish Symposium on Artificial Intelligence and Artificial Neural Networks*, Ankara, Turkey.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Philadelphia, USA, (10) pp. 79–86.

Seslisözlük Group. 2014. Sesli sözlük, bidirectional English-Turkish-English dictionary. http://www.seslisozluk.net, Accessed: September 2014.

Sureka, A., Goyal, V., Correa, D., and Mondal, A. 2012. Polarity classification of subjective words using common-sense knowledge-base. *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*. Berlin Heidelberg: Springer, pp. 486–93.

Thelwall, M., Buckley, K., and Paltoglou, G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* **63**(1): 163–73.

Turkmenoğlu, C., and Tantuğ, A. C. 2014. Sentiment analysis in Turkish media. In *Proceedings of Workshop on Issues of Sentiment Discovery and Opinion Mining, International Conference on Machine Learning (ICML)*, Beijing, China.

Vural, A. G., Cambazoğlu, B. B., Şenkul, P., and Tokgöz, Z., Ö. 2013. A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish. In *Computer and Information Sciences*, pp. 437–45. London: Springer.

Wang, Q., Cambria, E., Liu, C., and Hussain, A. 2013. Common sense knowledge for handwritten Chinese text recognition. *Cognitive Computation*, Springer **5**(2): 234–42.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Vancouver, B.C., Canada, pp. 347–54.

Wilson, T., Wiebe, J., and Hoffmann, P. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, **35**(3): 399–433.

Yıldırım, E., Çetin, F., S, Eryiğit, G., and Temel T. 2015. The impact of NLP on Turkish sentiment analysis. In *Proceedings of the TURKLANG'14 International Conference on Turkic Language Processing*, Istanbul, pp. 6–7.

https://doi.org/10.1017/S1351324916000309 Published online by Cambridge University Press