Sentiment Analysis in Turkish

Chapter · July 2018

DOI: 10.1007/978-3-319-90165-7_12

2 authors:

Gizem Gezici
Sabanci University

Berrin A. Yanikoglu
Sabanci University

Chapter 1
Sentiment Analysis in Turkish

Gizem Gezici and Berrin Yanıkoğlu

Abstract In this chapter, we give an overview of sentiment analysis problem and present a system to estimate the sentiment of movie reviews in Turkish. Our approach combines supervised learning and lexicon-based approaches and making use of a recently constructed Turkish polarity lexicon called SentiTurkNet. For performance evaluation, we investigate the contribution of different feature sets, as well as the effect of lexicon size on the overall classification performance.

1.1 Introduction

Sentiment analysis aims to identify the *polarity* and strength of the opinions indicated in a given text, that together define its *semantic orientation*. The polarity can be indicated categorically as *positive*, *objective* or *negative*, or numerically, indicating the strength of the opinion on a canonical scale.

Automatic extraction of the sentiment can be very useful in analyzing what people think about specific issues or items, by analyzing large collections of textual data sources such as personal blogs, product review sites, and social media. Commercial interest to this problem has been strong, with companies showing interest to public opinion about their products and financial companies offering advice on the general economic trend by following the sentiment in social media (Pang and Lee, 2008). In the remainder of this chapter, we use the terms "document", "review" and "text" interchangeably, to refer to a text whose sentiment polarity or opinion strength is to be estimated.

Gizem Gezici
Sabancı University, Istanbul, Turkey, e-mail: gizemgezici@sabanciuniv.edu

Berrin Yanıkoğlu
Sabancı University, Istanbul, Turkey, e-mail: berrin@sabanciuniv.edu

Approaches

There exist two fundamental approaches for sentiment analysis in the state-of-the-art: (i) linguistic or lexicon-based (Turney, 2002) and (ii) statistical or based on *supervised learning* (Pang et al, 2002). The first approach has the advantage of being simple, while the second approach is typically more successful since it learns from samples of documents with known sentiment in the given domain, without necessarily relying on specially compiled lexicons.

A *polarity lexicon* contains the sentiment polarity of words or phrases. SentiWordnet (Esuli and Sebastiani, 2006) and SenticNet (Poria et al, 2012) are two of the most commonly used domain-independent polarity lexicons, for sentiment analysis. The lexicon-based approach obtains the polarities of the words or phrases in a document from a polarity lexicon, for the goal of determining the semantic orientation of the document (Turney, 2002). The approach may be as simple as estimating the document polarity from that of the average polarity of the constituent words, or can be more complex where different properties of the text can be exploited with the hope of obtaining a more accurate semantic orientation. For instance, the number of subjective words in the document, or the *purity* of the constituent words may be considered (Taboada et al, 2011). The distinctive aspect of lexicon-based approaches is that they do not involve any machine learning.

Supervised learning approaches, *learn from from data*. While different learning techniques vary on how they use the available labelled data, called the *training data*, the common approach represents each review in the training data using a vector of features (e.g., length, average word polarity, etc.), and a model is learned to associate a feature vector representation with the desired output. The problem can be approached as a *classification* problem (e.g., a review is classified as positive or negative) or *regression* problem (e.g., number of stars given by a review is predicted.) Furthermore, classification can be *binary* (positive/negative) or *ternary* (positive/negative/neutral). The problem gets more difficult as the number of classes increase.

The model that is learned using the training data is tested on a separate *test* data, in order to gauge the generalization performance of the system. However, if there is no designated test set, the available data is split into two datasets as *train* and *validation*, such that the success of the model trained using the training portion is evaluated using the validation portion. For evaluation, the estimated labels (class labels or regression values) are compared with the true labels assigned in the validation/test data.

In case the available data is not very large, instead of splitting the data as training and test, one can make use of a technique called *cross-validation*. Cross-validation is a model validation technique for assessing how well the results of a predictive model would be generalized to an independent dataset. In k-fold cross-validation, the whole data is divided into k equal-sized subsets, where $k-1$ subsets are used for training and one is used for testing. To reduce variability, this train-test cycle is repeated for multiple rounds and at each round different partitions are used for training and test, and the results obtained with the each test data are averaged at the end.

In the basic approach to sentiment analysis, a given text is viewed as a bag-of-words (BoW); that is, it is represented as a set of words ignoring any word order information (Pang et al, 2002). With this representation, the sentence “A is better than B” is the same as “B is better than A”. However, in many cases, the loss of word order information may not necessarily have such a drastic impact (e.g., in “excellent movie was”). Alternatives to the bag-of-words approach are also possible, where word polarities of sentences at significant locations (e.g., first and last sentences) are taken into consideration (Zhao et al, 2008; Gezici et al, 2012).

Some of the supervised learning methods require a polarity lexicon in addition to the training corpus, in order to extract features of the given text (e.g., average word polarity, length, and the number of negative words, etc.) that are later exploited in the learning algorithm. In the Latent Dirichlet Allocation (LDA) approach, one of the most successful supervised learning approaches, the probability distributions of topic and word occurrences in different categories (e.g., positive and negative reviews) are learned by using a training corpus and the classification of a new text is done based to its likelihood estimated from these different distributions (Bespalov et al, 2011, 2012). Deep learning approaches found to be successful in many different pattern recognition problems in recent years are also applied to sentiment analysis with good results (Severyn and Moschitti, ????; Tang et al, 2014; Socher et al, 2013).

Data Type

There are two main types of data for which automatic sentiment analysis is of interest. In *reviews*, the text is generally longer and writers express their opinions on different aspects of the product (e.g., a movie, a hotel, a cell phone). In contrast, with *tweets*, writers express opinions using a very short text which brings on additional complications during automatic processing.

In the hotel domain, the TripAdvisor dataset is well-known, consisting of reviews that are crawled from the TripAdvisor website.[1] In the movie domain, there is a dataset of reviews from IMDB website.[2] For product reviews, researchers often rely on online product reviews from various websites such as Amazon.[3] Tweets in various topics can be collected using keyword searches but are generally more difficult to analyze for sentiment, owing to their short length, prevalence of spelling error and/or abbreviations, though, there have been significant developments on this through a yearly evaluation campaigns organized at relevant conferences (Rosenthal et al, 2014).

[1] `www.tripadvisor.com`

[2] `www.imdb.com`

[3] `www.amazon.com`

Domain Dependence

Words may have different meanings in different domains. For instance, the word "small" has a negative connotation in the hotel domain whereas it is in general positive in the cellphone domain. Since domain-independent lexicons such as SentiWordNet (Esuli and Sebastiani, 2006) and SenticNet (Poria et al, 2012) do not contain homonyms (a word that has diverse meanings in different contexts), they may mislead the sentiment analysis system. Hence, one may need a domain-specific lexicon which can be constructed by using a corpus of labeled reviews in a specific domain.

In the rest of the chapter, after a brief discussion of the related work, we discuss difficulties encountered in Turkish sentiment analysis and describe a Turkish sentiment analysis framework and give experimental results on a movie dataset.

1.2 Related Work

Research in sentiment analysis has been active for the last fifteen plus years in line with increasing academic and commercial interest in the topic. Pang and Lee (2008) present a detailed survey of the previous work. Here we only summarize research about the fundamental issues in sentiment analysis and discuss issues in Turkish sentiment analysis and the described and evaluate a system for Turkish.

In their seminal work, Pang et al (2002) evaluate several features with three different machine learning methods, on a dataset collected from movie reviews crawled from the IMDB internet movie database. A Support Vector Machine (SVM) classifier taking as input the occurrence counts of unigrams and bigrams give the best performance (82.9% accuracy on 1400 movie reviews).

As mentioned earlier in earlier work a document was typically viewed as a bag-of-words and its sentiment orientation was estimated from features extracted from this "bag" (e.g., words and their frequencies, etc.) (Hatzivassiloglou and McKeown, 1997; Pang et al, 2002; Pang and Lee, 2004; Mao and Lebanon, 2006). As word-order information was no longer available, researchers explored different methods so that they can analyze phrases and sentences. Wilson et al (2004) represented the document data in a tree structure and then generated features for displaying the relations in the tree with the help of boosting and rule-based methods. Gezici et al (2012) analyzed sentence-level features in order to bridge the large gap between word and document-level sentiment analysis.

While exploring features at different levels, researchers observed that one of the most important review properties that is highly relevant to sentiment analysis is *subjectivity*. They found that identifying subjective parts within the text first, may help to estimate the overall sentiment more accurately. Wiebe (2000) investigated the impact of adjective orientation and gradability on sentence subjectivity. The aim of the approach is to understand whether a given sentence is subjective or not, by looking at the adjectives in that sentence. Wiebe et al (2004) presented a broad subjectivity analysis along with a comprehensive survey of subjectivity recognition using various features and clues. One of the first datasets generated for the classification of subjectivity consists of 5000 subjective movie review snippets and 5000 movie plot sentences which are assumed to be objective (Pang and Lee, 2004). Using this dataset, Pang and Lee built a two-layer algorithm for classification, where the first layer differentiated subjective sentences from objective ones and classified subjective sentences as positive or negative. The two-layer classification process increased the overall result by 4% from 82.9% accuracy (Pang et al, 2002) to 86.9% accuracy (Pang and Lee, 2004).

Since 2007, SemEval (Semantic Evaluation) evaluation campaigns have promoted and benchmarked progress in sentiment analysis, receiving a large participation from around the world. Best systems in SemEval 2015 consist of deep learning techniques and ensemble methods (Severyn and Moschitti, ????; Hagen et al, 2015) and achieve around 65% accuracy in ternary classification of tweets.

Most sentiment analysis research in literature is for English and most resources for sentiment analysis (e.g., polarity lexicons, parsers) are for English as well. However research on sentiment analysis of non-English texts has picked interest in recent years. For instance, Ghorbel and Jacot (2011) formulated a method to classify French movie reviews by using supervised learning and linguistic features that are extracted with part-of-speech tagging and chunking, using the semantic orientation information of words from SentiWordNet (Esuli and Sebastiani, 2006). French words in the reviews are translated to English so as to obtain their semantic orientation from SentiWordNet.

Sentiment analysis of texts in Turkish has also attracted research interest in recent years and there is still a lot to do in the field. Eroğul (2009) is one of the first studies on Turkish sentiment analysis and develops an SVM classifier for Turkish movie reviews, crawled from the web-size BeyazPerde.[4] It uses n-grams as features and studies the effect of part-of-speech tagging, spelling-checking and stemming on the overall result. The classifier achieves an 85% of accuracy on the binary sentiment classification of Turkish movie reviews.

More recently, Vural et al (2013) has proposed a lexicon-based sentiment analysis system using SentiStrength, a lexicon-based sentiment analysis library developed by Thelwall et al (2010). The library generates a positive or a negative sentiment score for each word in a given text. The authors evaluated their unsupervised Turkish sentiment analysis framework on the same dataset that was already used in Eroğul (2009) and reported an accuracy of 76% for positive/negative classification.

Türkmenoğlu and Tantuğ (2014) present a lexicon-based framework similar to the systems described in Thelwall et al (2010) and Vural et al (2013), with additional handling of simple negation and multi-word expressions. They report achieving 79.0% and 75.2% accuracy on movie review and tweets datasets, respectively.

As a related problem, *emotion analysis* has also attracted attention from researchers. Boynukalın (2012) presents an emotion classification framework for

[4] `www.beyazperde.com`

Turkish and comparative experimental results of several classifiers, for a new dataset for Turkish emotion analysis, with an investigation of the effects of newly added features which are compatible with the morphological characteristics of Turkish language. In another study on emotion analysis of Turkish texts, Çakmak et al (2012) investigates the feasibility of using a fuzzy-logic representation of Turkish emotion-related words and indicates that there is a strong connection between emotions expressed by word roots and sentences in Turkish.

Analysis of Turkish political news and tweets has also attracted researcher interest. Kaya et al (2012) investigate the performance of supervised machine learning algorithms such as Naive Bayes, maximum entropy, SVM and the character based n-gram language models. They observe that maximum entropy and the n-gram models outperform the SVM and Naive Bayes approaches, and report 76-77% accuracy using different features. Kaya (2013) describes an improved version of this system by implementing transfer learning into the existing framework. This system accomplishes a significant improvement over the previous systems, and with all of the three machine learning approaches above, reports accuracy values over 90% for the sentiment classification of Turkish political opinion columns.

1.3 Main Difficulties for Turkish Sentiment Analysis

The motivation behind building a sentiment analysis framework specific to Turkish, rather than utilizing an already established system for English and translating it into Turkish, is due to certain important differences between these two languages. These main differences can be summarized in three categories as follows:

1. *Morphology of Turkish*: Turkish is an agglutinative language (see Chap. **??**) which allows the generation of many inflected and derived variant forms of a word adding suffixes to a root word. These derivational and inflectional suffixes may change the part-of-speech and semantic orientation of the word (e.g., “beğendim” (I liked it), “beğen<u>me</u>dim” (I didn’t like it)).

   The practical effect of the agglutinative morphology in sentiment analysis is that it makes it infeasible to build a (polarity) lexicon that would contain all variants of Turkish words. Hence, sentiment analysis systems for agglutinative languages like Turkish face some extra challenges compared to those for which a reasonable size lexicon (e.g., 30,000 as for English) is sufficient for many applications.

2. *Turkish character set*: Turkish has several characters that do not exist in the English alphabet: “ç”, “ğ”, “ı”, “ö”, “ş”, “ü”. In informal writing, people tend to substitute the closest ASCII characters for some of these letters (e.g., “ç" is written as “c"). While as such these words are readable in context by humans, they cause complications for robust determination of the words in a sentence.

A preprocessing stepknown as “deasciification” (i.e., converting the ASCII English characters to their Turkish equivalents) to find the words and obtain their polarities from the lexicon), is thus needed before sentiment analysis so that the input is in proper Turkish.

3. *Complexity of negation*: In Turkish, there are several ways a word may be negated, which in turn changes the expressed sentiment:
   - The suffixes *me/ma* negate a verb (e.g., “beğenmedim” (I did not like it)).
   - The productive derivational suffixes *siz/sız/suz/süz* derive an adjective whose meaning adds “’without” to the noun they are attached to (e.g., “başarısız” (without success – unsuccessful).
   - The word “değil” (is/are not) negates nominal or adjectival predications (e.g., “güzel değil” (is not beautiful))
   - The word“yok” (does not exist) indicates nonexistence or unavailability (e.g., “konusu yok” (does not have a topic))

1.4 Practical Sentiment Analysis for Turkish

In this section, we present a basic sentiment analysis system for Turkish, starting with a basic baseline approach and showing how subsequent steps utilizing simple natural language processing (NLP) steps and increasing the lexicon size improve the basic results.

Our evaluation procedure is composed of two main parts: first, we report the effectiveness of different sets of features in clsssifying movie reviews in Turkish. We then investigate the influence of lexicon size on detecting the overall sentiment of the reviews in the same dataset.

1.4.1 Resources

1.4.1.1 Polarity Lexicon

Our polarity lexicon is the first comprehensive Turkish polarity lexicon, *SentiTurkNet*, described in Dehkharghani et al (2015) and developed using several resources both in English as well as in Turkish. In building this lexicon, authors did not translate SentiWordNet to Turkish as was done in Türkmenoğlu and Tantuğ (2014), but rather they compiled the lexicon using various NLP techniques and additional available resources such as Turkish WordNet (Bilgin et al, 2004), English WordNet (Fellbaum, 1998), SentiWordNet (Esuli and Sebastiani, 2006) and SenticNet (Cambria et al, 2010).

SentiTurkNet consists of 15,000 synsets with their part-of-speech tags[5] and three associated polarity values – positive, negative and neutral/objective. The polarity scores stand for the measurement of negativity, objectivity and positivity, and sum up to 1. Some sample entries from SentiTurkNet are provided in Table 1.1.

Note that a given word may belong to different synsets with different sentiment polarity and hence, one needs to find the correct synset to obtain the correct sentiment polarity. If this is not feasible, then sentiment polarity values accross different synsets corresponding to the given word may be averaged. We took the latter approach in this work.

Table 1.1 Sample Entries from SentiTurkNet

Synset	POS-Tag	Negative	Objective	Positive
mükemmel, kusursuz (excellent)	a	0.000	0.000	1.100
kötü (bad)	a	0.946	0.018	0.036
çekici, güzel (beautiful, attractive)	a	0.000	0.000	1.000
şaka, latife (joke)	n	0.060	0.397	0.543
gülmek (to laugh)	v	0.095	0.095	0.810
fiilen, gerçekten (really)	b	0.060	0.872	0.068

1.4.1.2 Seed Words

Seed words are highly sentiment-bearing words (e.g., “excellent”, “horrible”) that are expected to be strong indicators of a review’s sentiment. Seed word sets have been commonly used for sentiment analysis (Hu and Liu, 2004; Qiu et al, 2011).

Another advantage of sentiment-bearing words is that their sentiment polarity does not change much across different domains. For instance, *muhteşem* (excellent) is a positive word in Turkish and *berbat* (awful) is a negative word, independent of the context. Hence they may be helpful in domain-independent tasks.

Yet another important quality of seed words is that they are often not used in negated form, simplifying the analysis of the sentiment they carry. For instance, while one may use “not very good”, where the sentiment polarity of the word “good” is reversed, it is not common to use highly sentiment-bearing words in negated form (as in “it was not excellent”).

We use a positive seed word list of 34 words and a negative seed word list of 93 words in this work. A sample of 10 positive and 10 negative seed words from this list are shown in Table 1.2.

1.4.1.3 Booster Word List

Booster words are adverbs that accentuate the sentiment polarity of the words that follow: For instance, the words “very” or “really”, as in “it was a really good movie”,

[5] We label these as follows in this chapter: a – adjective, n– noun, v–verb, and b–adverb.

Table 1.2 Sample Seed Words

Positive Words	Type	Negative Words	Type
muhteşem (magnificient)	a	fiyasko (failure)	n
güzel (beautiful)	a	berbat (awful)	a
eğlenceli (enjoyable)	a	hayalkırıklığı (disappointment)	n
harika (awesome)	a	sıradan (average)	a
şahane (fantastic)	a	sıkıcı (boring)	a
etkileyici (fascinating)	a	olumsuz (negative)	a
başyapıt (masterpiece)	n	vasat (mediocre)	a
kaliteli (good quality)	a	felaket (terrible)	a
kusursuz (perfect)	a	beğenmedim (I did not like)	v
inanılmaz (incredible)	a	değmez (not worth it)	v

are examples of such words. The boosting effect has already been investigated for Turkish (Türkmenoğlu and Tantuğ, 2014).

We have a very small list of four commonly used booster words shown in Table 1.3. Strengthening is done by shifting the polarity value of the corresponding adjective towards its sentiment pole, i.e., positive or negative. We chose a value of 0.4 for shifting.

Table 1.3 Booster Words

Word	POS Tag
en (most)	b
gerçekten (really)	b
çok (very)	b
bayağı (too many/much)	b

1.4.2 Methodology

Our approach combines supervised learning and lexicon-based approaches. In the baseline approach, we simply compute the average polarity of the words (adjectives, verbs, and nouns) in the review and train a classifier (Naive Bayes or SVM) to classify the reviews as positive or negative just based on this average polarity feature. Then we measure the effectiveness of more complex processing techniques or additional features such as handling negation, considering the effects of booster words and using additional features derived from seed words. In all of these approaches, the document is viewed as a bag-of-words.

1.4.2.1 Preprocessing

Before feature extraction, several steps are necessary as preprocessing steps, in order to obtain the corresponding polarity values of the word in a review. These polarity values form the basis of the features used in this work. As an initial step, we tokenize the given text into words and then we use Zemberek (Akın and Akın, 2007) for deasciification.

While obtaining the polarity value for each word in the document, we follow the following procedure: We first search the word itself in the lexicon (SentiTurkNet) with the part-of-speech tag information. If the word is not found, then we identify the root with Zemberek and search for the root in the lexicon. If we still cannot find it and the part-of-speech tag of the word is verb, then we search the root of the word by adding the infinitive suffixes (mek/mak in Turkish) to the end of it. If none of these help to find the polarity values for the word, this means that the word does not exist in the lexicon, therefore its polarity values are set to 0. The process is illustrated in Table 1.4 for the sample word “hoslanmadim (I did not like it)", written using only ASCII versions of some of the characters.

Table 1.4 Sample Preprocessing

Input	Preprocessing Step	Output	Lexicon Search
hoslanmadim	Deasciification	hoşlanmadım	Not Successful
hoşlanmadım	Root extraction	hoşlan	Not Successful
hoşlan	Adding infinitive suffixes	hoşlanmak	Successful

A word in the lexicon may have multiple synset entries. In order to get the correct polarity values, it is important to find the correct synset, or as a lesser alternative, to compute an average of the polarity values of all corresponding synsets. We take the latter approach in this work for simplicity.

1.4.2.2 Basic approach

In the basic approach, we only use the average polarity of the constituent words to estimate the document polarity. The overall average sentiment polarity is computed by averaging the polarity of all potentially sentiment-bearing words in the document (adjectives, verbs and nouns), while adverbs affect the overall polarity indirectly if they are in the booster list shown in Table 1.3. The average polarity of a given text is computed as follows:

$$F_1 = \frac{1}{N} \sum_{w_i} pol(w_i) \tag{1.1}$$

where w_i are the corresponding words in the document, N is the total number of sentiment-bearing words and $pol(w_i)$ is computed from the polarity values obtained from SentiTurkNet. The *average polarity* of a word w, denoted by $pol(w)$, is calculated as:

$$pol(w) = (pol^{+} - pol^{-})/2 \tag{1.2}$$

where pol^{+} and pol^{-} represent the positive and negative polarity values assigned to the word w in the polarity lexicon. For simplicity, we do not take into account of the neutral/objective polarity value of the word in this work. An alternative to using the average polarity is to use the *dominant polarity* of a word (Demiröz et al, 2012).

1.4.2.3 Handling negation

In comparison to English, negation handling is quite complicated for Turkish. For instance, in English the word *not* is used for negation purposes, while in Turkish negation can happen in several different forms as described earlier. We take into consideration all words or suffixes that signal negation except *yok* because it requires the negation analysis at the sentence instead of word-level.

For each negated word, we negate its polarity $pol(w_i)$ as defined in Eq. 1.2 and recompute the average polarity to give feature F_2.

1.4.2.4 Booster Effect

As mentioned in earlier booster words strengthen the meaning of adjectives that they modify. In order to take them into account, we compute the average review polarity by considering the effects of booster words shown in Table 1.3, by shifting the the polarity of affected words. Booster word handling is performed after negation handling, to obtain feature F_3.

to obtain the feature F_3.

1.4.2.5 Seed Words

We have chosen a positive seed word list of 34 words and a negative seed word list of 93 words, as discussed earlier. The seed word enables us to estimate sentiment that is less error-prone in comparison to using a large polarity lexicon that may contain errors. The corresponding features F_4 and F_5 are the positive and negative seed word counts in a review. Feature F_4 is computed as:

$$F_4 = \sum_{w_i} PositiveSeed(w_i)$$

where w_i are the sentiment-bearing words that are adjectives, adverbs, verbs and nouns in the document and $PositiveSeed(w_i)$ returns 1 if the word w_i is a positive seed word and zero otherwise. Similarly, feature F_5 captures the number of negative seed words in the review.

1.4.2.6 Sample Analysis

Table 1.5 shows the feature values for three separate sentences. The first example has a single sentiment bearing word, no booster words nor any negation suffix. Hence all three features have the same value. In the second example there is a negation suffix (-me), which is considered in determining the average polarity in F_2. As there are no booster words, F_3 is the same as F_2. The third example contains a booster word si therefore the average polarity of the following adjective is shifted by 0.4 towards the positive end in determining F_3.

Table 1.5 Sample Sentences

Sample Input	Relevant Words	F_1	F_2	F_3
Hata-larla dolu (full of errors)	hata (n; pol=-0.47)	-0.47	-0.47	-0.47
Hiç sev-me-dim (I did not like at all)	sev (v; pol=0.37), -me (negation)	0.37	-0.37	-0.37
Cok guzel-di (was very beautiful)	çok (booster), güzel (a; pol=0.5)	0.50	0.50	0.90

1.4.2.7 Classifier Training

We randomly split the available data into train and test sets containing a balanced number of positive and negative reviews in each. Then, the system is trained using a Naive Bayes or SVM classifier using only the training set and tested on the test set. Our system is implemented in Java and uses WEKA(Hall et al, 2009) for classifier training and testing. WEKA is a commonly used machine learning toolbox that provides many supervised as well unsupervised algorithms (Hall et al, 2009).

We used the LibSVM package which is implemented in WEKA for parameter optimization, training and testing stages. Before the actual training with the SVM classifier, we performed parameter optimization.7 For optimization, we performed a 5-fold cross-validation on the training data and found the best parameter values of 10.0 and 10.0 for the cost and gamma parameters. We then re-trained the system with all of the training data.

1.5 Experimental Evaluation

1.5.1 Data

We evaluated the proposed approach and features using the Turkish movie reviews dataset that was compiled by Demirtaş and Pechenizkiy (2013) from a well-known movie site called Beyazperde.[6] The star ratings of reviews (1 to 5 stars) are used as ground-truth labels for evaluation. Since we only address the binary classification problem, 4 or 5-star reviews are considered as positive reviews while 1 or 2-star reviews are considered negative. We excluded reviews with 3-stars from the study, as often done in binary classification evaluations. As a result, we obtained a total set of 5331 positive and 5330 negative movie reviews. Some sample positive and negative movie reviews from the database are shown in Table 1.6 and 1.7, respectively. The data is split into train and test sets with equal proportion of positive and negative reviews in each.

Table 1.6 Positive Movie Reviews

Review	Gloss
"gerçek bir başyapıt"	" a true masterpiece"
"gelmiş geçmiş en iyi 10 filmden biri"	"it's one of the top 10 movies ever"
"tek kelimeyle kusursuz"	"in one word: perfect"

Table 1.7 Negative Movie Reviews

Review	Gloss
"benim için sadece büyük bir hayalkırıklığı"	"for me it's just a big disappointment"
"hiç beğenmedim bu filmi"	"I didn't like this movie at all"
"berbat bir film"	"it's a terrible movie"

1.5.2 Results

We report results obtained with both Naive Bayes or SVM classifiers, using the features in increasing complexity. Table 1.8 presents correct classification accuracies with basic and more complex features, while Table 1.9 presents the effect of the lexicon size in overall accuracy.

[6] Reviewers on Beyazperde rate movies star ratings of 1 to 5 scale, in addition to the review they enter.

Feature Efficacy

While the basic baseline approach only uses the raw sentiment polarities of words in estimating the average polarity of a given document, the second approach extends it with negation handling to compute the average document polarity more accurately, and the third approach includes both negation and booster word handling. As we see in Table 1.8, the basic approach obtains 67.49% accuracy with the Naive Bayes classifier and 67.61% with the SVM classifier, while best results are obtained with negation handling and seed words, achieving 75.16% with the Naive Bayes and 73.70% with the SVM classifiers, respectively. Somewhat surprisingly, booster effect handling does not improve classification accuracy significantly, while considering seed words does.

Table 1.8 Classification Accuracy with Different Features

Features	Accuracy (NB)	Accuracy (SVM)
F_1 (Basic)	67.49%	67.61%
F_2 (w/ neg. handling)	69.29%	69.42%
F_3 (w/ neg. + booster handling)	68.22%	68.19%
F_2, F_4, F_5 (neg. handling + seed words)	75.16%	73.70%

Lexicon Effect

The second part of our evaluation investigates the effect of the lexicon size on obtaining the overall sentiment of a given review. Increasing the lexicon size generally improves the classification performance, since with a larger lexicon, the system knows about the semantic orientations of more words.

To generate lexicons of various sizes, we started with the polarity values of the seed words, obtained from the SentiTurkNet (Dehkharghani et al, 2015), and the rest of the new lexicon was filled by randomly choosing the necessary number of synsets from the lexicon. To obtain more robust results, we randomly chose the rest of the words in the new lexicon five times and obtained results and computed an average over these.

This process was repeated until the lexicon size reached that of SentiTurkNet which contains 15000 synsets. In investigating the effect of lexicon size, we only used our basic feature, F_1. Results are displayed in Table 1.9 where the last row corresponds to the basic approach given in Table 1.8. As can be seen, a larger lexicon always brings better classification performance as the added words help estimate the review polarity more accurately.

Table 1.9 The Effects of Lexicon Size on the Classification Performance

Lexicon Size	Accuracy (NB)	Accuracy (SVM)
100	51.27%	51.29%
1000	51.85%	51.88%
5000	52.07%	53.28%
(All) 15000	67.49%	67.61%

1.6 Conclusions

Interest in sentiment analysis is growing rapidly thanks to its use in collecting public's opinion in several different application areas. Various approaches described in literature range from simple approaches based on the use of domain-independent polarity lexicons to deep learning techniques that can capture long-term interactions between words in a review. Suitability and success of different approaches depend upon many factors, including the availability polarity resources in the given language and domain, availability of labeled data, the length of the review to be analyzed. Through a simple system, we have demonstrated the effects of some of the necessary natural language processing steps (i.e. negation handling and booster word handling), along with the effect of seed words and lexicon size. We achieve 75% accuracy on binary classification of movie reviews in Turkish. Our results show that having even a small set of domain-dependent seed words and a large domain-independent polarity lexicon affects recognition accuracy the most. Future work in this area seems to be geared towards building resources in new languages, as well as machine learning techniques such as deep learning that leverage large amounts of unlabeled data, in addition to labeled data and sentiment resources.

References

Akın AA, Akın MD (2007) Zemberek, an open source NLP framework for Turkic languages. Structure 10

Çakmak O, Kazemzadeh A, Yıldırım S, Narayanan S (2012) Using interval type-2 fuzzy logic to analyze Turkish emotion words. In: Signal Information Processing Association Annual Summit and Conference, Los Angeles, CA, pp 1–4

Bespalov D, Bai B, Qi Y, Shokoufandeh A (2011) Sentiment classification based on supervised latent n-gram analysis. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, Glasgow, Scotland, pp 375–382

Bespalov D, Qi Y, Bai B, Shokoufandeh A (2012) Sentiment classification with supervised sequence embedding. In: Machine Learning and Knowledge Discovery in Databases, Springer, pp 159–174

Bilgin O, Çetinoğlu Ö, Oflazer K (2004) Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology 7(1-2):163–172

Boynukalın Z (2012) Emotion analysis of Turkish texts by using machine learning methods. Master's thesis, Middle East Technical University, Ankara, Turkey

Cambria E, Speer R, Havasi C, Hussain A (2010) Senticnet: A publicly available semantic resource for opinion mining. In: AAAI Fall Symposium: Commonsense Knowledge, Arlington, VA, vol 10, p 02

Dehkharghani R, Saygın Y, Yanıkoğlu B, Oflazer K (2015) Sentiturknet: A Turkish polarity lexicon for sentiment analysis. Language Resources and Evaluation

Demiröz G, Yanıkoğlu B, Tapucu D, Saygın Y (2012) Learning domain-specific polarity lexicons. In: SENTIRE: Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, Brussels, Belgium, pp 674–679

Demirtaş E, Pechenizkiy M (2013) Cross-lingual polarity detection with machine translation. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '13, pp 9:1–9:8

Eroğul U (2009) Sentiment analysis in Turkish. Master's thesis, Middle East Technical University, Ankara, Turkey

Esuli A, Sebastiani F (2006) Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, vol 6, pp 417–422

Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA

Gezici G, Yanıkoğlu B, Tapucu D, Saygın Y (2012) New features for sentiment analysis: Do sentences matter? In: The 1st International Workshop on Sentiment Discovery from Affective Data, Bristol, UK, pp 5–15

Ghorbel H, Jacot D (2011) Sentiment analysis of French movie reviews. In: Advances in Distributed Agent-Based Retrieval Tools, Springer, pp 97–108

Hagen M, Potthast M, Büchner M, Stein B (2015) Webis: An Ensemble for Twitter Sentiment Detection. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 15), Association for Computational Linguistics, pp 582–589

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11(1):10–18

Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, Spain, pp 174–181

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Seattle, WA, pp 168–177

Kaya M (2013) Sentiment analysis of Turkish political columns with transfer learning. PhD thesis, Middle East Technical University, Ankara, Turkey

Kaya M, Fidan G, Toroslu İH (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, IEEE Computer Society, Macau, pp 174–180

Mao Y, Lebanon G (2006) Isotonic conditional random fields and local sentiment flow. In: Advances in Neural Information Processing Systems, pp 961–968

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Barcelona, Spain, p 271

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, pp 79–86

Poria S, Gelbukh A, Cambria E, Das D, Bandyopadhyay S (2012) Enriching SenticNet polarity scores through semi-supervised fuzzy clustering. In: SENTIRE: Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE, Brussels, Belgium, pp 709–716

Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Computational Linguistics 37(1):9–27

Rosenthal S, Ritter A, Nakov P, Stoyanov V (2014) Semeval-2014 task 9: Sentiment analysis in twitter. In: Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, pp 73–80

Severyn A, Moschitti A (????) Unitn: Training deep convolutional neural network for twitter sentiment classification. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 15), Association for Computational Linguistics, pp 464–469

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, pp 1631–1642

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2):267–307

Tang D, Wei F, Qin B, Liu T, Zhou M (2014) Coooolll: A deep learning system for twitter sentiment classification. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp 208–212, URL `http://www.aclweb.org/anthology/S14-2033`

Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12):2544–2558

Türkmenoğlu C, Tantuğ AC (2014) Sentiment analysis in Turkish media. Tech. rep., Istanbul Technical University

Turney PD (2002) Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, PA, pp 417–424

Vural AG, Cambazoğlu BB, Şenkul P, Tokgöz ZÖ (2013) A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. In: Computer and Information Sciences, Springer, pp 437–445

Wiebe J (2000) Learning subjective adjectives from corpora. In: Proceedings of AAAI, Austin, TX, pp 735–740

Wiebe J, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Computational Linguistics 30(3):277–308

Wilson T, Wiebe J, Hwa R (2004) Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI, San Jose, CA, pp 761–769

Zhao J, Liu K, Wang G (2008) Adding redundant features for CRF-based sentence sentiment classification. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, pp 117–126