Building Phrase Polarity Lexicons for Sentiment Analysis

Rahim Dehkharghani*

Faculty of Engineering, University of Bonab, Bonab, I.R. (Iran)

Received 11 May 2018 | Accepted 17 October 2018 | Published 19 October 2018

Abstract

Many approaches to sentiment analysis benefit from polarity lexicons. Most polarity lexicons include a list of polar (positive/negative) words, and sentiment analysis systems attempt to capture the occurrence of those words in text using polarity lexicons. Although there exist some polarity lexicons in many natural languages, most languages suffer from the lack of phrase polarity lexicons. Phrases play an important role in sentiment analysis because the polarity of a phrase cannot always be estimated based on the polarity of its parts. In this work, a hybrid approach is proposed for building phrase polarity lexicons which is experimented on Turkish as a low-resource language. The obtained classification accuracies in extracting and classifying phrases as positive, negative, or neutral, approve the effectiveness of the proposed methodology.

Keywords

Sentiment Analysis, Polarity Lexicons, Polarity Classification, Phrases.

DOI: 10.9781/ijimai.2018.10.004

I. Introduction

Due to ever-increasing amount of online information especially in social media, manual processing of data to extract valuable information is impractical. The task of extracting information from text might attempt to extract the polarity of text--which is called sentiment analysis or polarity classification. This task has been very popular in recent decades but still it is far from the ideal.

Many approaches to sentiment analysis require polarity lexicons to assign a polarity tag (positive, negative or neutral) to a segment of text. There exist a good deal of workA on polarity lexicon generation which is grouped into two categories by Liu [1]: dictionary based methods and corpus based methods. Dictionary based methods start with a small seed word list and expand it upon synonymy and antonymy relations by using dictionaries such as WordNet [2]. In corpus based methods, semantic relations between terms in a corpus are employed to generate polar terms. These relations include pointwise mutual information [3] considering the co-occurrence of words in a window (e.g., a sentence), conjoined adjectives (by ``and'' or ``but'') [4], and delta tf-idf [5].

In this paper, a novel approach has been suggested for generating and classifying phrases as positive, negative, or neutral. The proposed approach is illustrated as a flowchart in Fig. 1. At first, raw phrases are collected; then, classification features are extracted; and finally, different classification tasks are accomplished to classify phrases as positive, negative, or neutral (objective). The contribution of this work is proposing a novel approach for generating phrase polarity lexicons and building the first phrase lexicon for the Turkish language. Note that the proposed approach is language independent, and it has been applied on Turkish as a case study. An alternative method for building such lexicons would be manually annotating the whole lexicon which has been employed in [6].

* Corresponding author.

E-mail address: rdehkharghani@bonabu.ac.ir

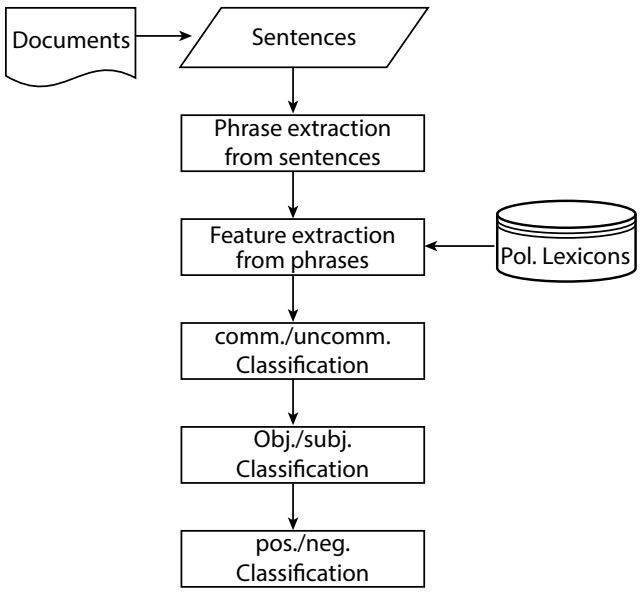

Fig. 1. The proposed methodology as a flowchart.

Among natural languages, most researchers have focused on English, while many other languages such as Turkish suffer from a lack of polarity resources. We have already generated two polarity lexicons for Turkish--Polar word Set (PWS) and SentiTurkNet (STN) in previous work [7]; however, because these lexicons have a limited coverage, a new polarity lexicon is generated in this work. Phrases require special attention in sentiment analysis because in most cases, the overall polarity of a phrase differs from the polarity of its parts. For example the phrase "...daha fazla olmalıydı" [...it should be more (better) than this] has a negative polarity but the constituting words are neutral. It seems that it is impossible to estimate negative polarity of this phrase based on the polarity of its parts.

In the remainder of this paper, Section II reports some previous works on sentiment analysis. The detailed explanation of the proposed approach is provided in Section III which is followed by experimental evaluation in Section IV. Applying the proposed method on other languages is discussed in Section V. Discussion on results are presented in Section VI, and Conclusions and future works are provided in Section VII.

II. Literature Review

Sentiment analysis can be done in different granularity levels: document, sentence, phrase, concept, and word levels. In [8], the authors investigated document level sentiment analysis using machine learning techniques.

In sentence level, Meena and Prabhakar [9] addressed the effect of conjunctions, and semantic relations between sentences.

In phrase level sentiment analysis, two works have been accomplished by Wilson and colleagues: [10] and [11]. In 2005, the authors proposed an approach which first classifies an expression as subjective or objective and then estimates its polarity in the case of subjectivity.

This method estimates the contextual polarity of an expression by using a large number of subjectivity clues and the prior polarity of appeared words in the expression. This work mostly relies on statistical methods. The obtained accuracies in classifying expressions as objective/subjective and also positive/negative range from 61% to 75%. The authors extended their work in 2009. The focus of this work is to figure out which features are more important in automatically distinguishing between prior and contextual polarity. Multi-perspective Question Answering (MPQA) is used as the opinion lexicon in this work.

Again in phrase level, Agrawal et al. [12] proposed a method to predict contextual polarity of subjective phrases in a sentence. The authors present new classification features which could achieve higher accuracies in ternary (positive/negative/neutral) classification of phrases over two baselines--majority class baseline as well as a more difficult baseline consisting of lexical n-grams.

Yi et al. [13] analyzed grammatical sentence structures and phrases for sentiment analysis purposes. The authors present Sentiment Analyzer which extracts sentiment towards a subject from online text documents. Instead of classifying the sentiment of an entire document about a subject, the designed system detects all references to the given subject, and determines the sentiment in each of the references.

In [14], the authors proposed an approach for extracting sentiments associated with positive or negative polarity for specific subjects in a document, instead of classifying the whole document as positive or negative. In this work, the goal is to identify semantic relationships between sentiment expressions and subject terms. Finally Kiritchenko and Saif [6] investigate phrases with opposite polarity such as *happy accident*. Phrases in this work are extracted from a large set of tweets using some patterns and they have been manually annotated by positive/negative tags.

In concept-level, Tsai et al. [16] presented a two-step methodology which combines iterative regression and random walk with in-link normalization to build a concept-level sentiment lexicon. In [16], the authors presented a methodology for enriching SenticNet [17]--a polarity lexicon in English-- concepts with affective information by assigning an emotion label to those concepts.

There exist also a good deal of research on building polarity lexicons. Liu [1] categorizes these methods into two groups: dictionary based approaches and corpus based approaches.

Dictionary based approaches start with a small seed set (e.g., 20 terms) and expand the list by using the existing relations such as synonymy and antonymy among terms in dictionaries. In [18], Hu and Liu used this method to generate a list of polar English terms and then manually cleaned up the generated list to remove errors. A similar approach was proposed in [19], which assigns also a sentiment score to each word by using a probabilistic method. Esuli and Sebastiani [20] proposed a methodology to assign three polarity scores (positive, negative, and neutral) to each synset in English WordNet. This approach was modified in [7] to build a polarity lexicon for Turkish based on the Turkish WordNet [21].

In corpus based approaches, having a seed word list with known polarities, new polar words are extracted based on the existing semantic relations in the corpus. One of the early ideas was proposed in [4]. The authors used conjunctions in a corpus to find new polar adjectives. They show that conjoined adjectives by "and" usually have the same polarity while they usually have the opposite polarity when conjoined by ``but''. Some extra relations such as "Either-or" and "Neither-nor" were also used for this purpose. Kanayama and Nasukawa [22] followed this approach and improved it by adding this idea: consecutive sentences usually have the same polarity.

There are also some effort on sentiment analysis of Turkish text. Yıldırım et al. [23] accomplished a sentiment analysis task on Turkish tweets in the telecommunication domain. The authors applied a multi-class ternary (positive/negative/neutral) classification by support vector machines on tweets using features such as inverse document frequency, unigrams and adjectives. They also benefit from NLP techniques such as normalization, stemming, and negation handling. Vural et al. [24] presented a system for unsupervised sentiment analysis in Turkish text documents using SentiStrength [25] 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. Kaya et al. [26] investigated sentiment analysis of Turkish political news in online media. The authors used four different classifiers--Naive Bayes, Maximum Entropy, SVM, and the 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 Maximum Entropy and the n-gram language models are more effective than the SVM and Naive Bayes classifiers in classifying Turkish political news. Boynukalın [27] has worked on emotion analysis of Turkish texts by using machine learning methods. The author has investigated four types of emotions: joy, sadness, fear, and anger on a dataset that she built for this purpose.

III. Phrase Polarity Lexicon Generation

A hybrid approach has been used for building a phrase polarity lexicon. The first phase in this approach is pre-processing. This pre-processing step as well as the whole approach are explained in the following subsections.

A. Phrase Extraction

A phrase is defined as "a small group of words standing together as a conceptual unit, typically forming a component of a clause" in Oxford dictionary[1]. As another definition from a Turkish dictionary[2], phrase is defined as "birkaç sözcükten oluşan ifade" (an expression composed of several words). According to Oxford dictionary, phrases can be divided into noun, verb, adjective, adverbial, and prepositional phrases[3]; however, only adjective, noun, and verb phrases are the focus of this work. According to Oxford dictionary, a noun phrase is a word or group of words containing a noun and functioning in a sentence as subject, object, or prepositional object such as "*inanılmaz bir*

[1] http://www.oxforddictionaries.com

[2] https://www.seslisozluk.net/

[3] https://en.oxforddictionaries.com/grammar/phrases

performans" (an unbelievable performance); A verb phrase is a verb with another word or words indicating tense, mood, or person such as "*gözlerimizi boyadılar*" (they deceived us); An adjective phrase is a phrase whose head is an adjective such as ``nasıl böyle saçma" (how silly like this).

At the first phase of the suggested methodology, a phrase list is generated by extracting collocations--trigrams and quadrigrams--using patterns in Table I, from 270,000 sentences in Turkish movie reviews (detailed explanation of the movie dataset is provided in Section IV.A). In this table, numbers inside parentheses are the number of phrases extracted by each pattern; moreover, one sample phrase has been provided for each pattern. The employed patterns (trigrams and quadrigrams) could extract 5213 phrases which are generally meaningful in Turkish; however, bigrams and 5-grams could extract more phrases, which is left as future work. Fig. 2 illustrates the percentage of each part of speech (POS) in extracted patterns. As seen in this figure, adjectives play the most important role and verbs play the least important role among other parts of speech. In this figure, number *P* upon a POS tag bar means that *P*% of phrases includes word(s) with the mentioned POS tag. In order to extract phrasal expressions from text, different methods could be used. One is exploited in this work, which is evaluated in Section IV. Extracting related words together in dependency parse tree is another method which was experimented in [28], but using those kind of phrases--which are usually separated by other words—for sentiment analysis purposes would be very challenging.

TABLE I. Patterns Used for Extracting Phrases from Sentences. Numbers Inside Parentheses Are the Number of Phrases Extracted by each Pattern

Triples	quadruples
adv adj verb (750) çok güzel anlatıyor	adv adj adj noun (436) çok iyi bir şekilde
adv adj noun (972) çok eğlenceli vakit	adv adj noun noun (394) bir güzel oyuncu hikayesi
adj noun verb (676) iyi iş çıkarmış	adj adj noun verb (491) abartılacak bir şey yoktu
adv adv verb (306) çok çok beğendim	adv adv adj verb (525) Neyse çok iyi diyemem
Adv adv adj (310) çok çok sevimli	adv adj noun verb (371) çok büyük saygı duyuyorum

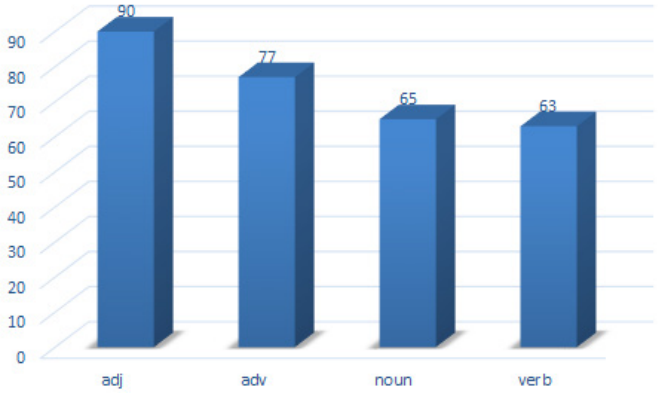

Fig. 2. The contribution of each POS tag in generation of phrases.

At this point a question might raise in mind that how well the suggested patterns can extract the existing phrases in the text. To answer this question, we manually extracted all correctly formed phrases from 100 randomly chosen sentences in Turkish movie reviews and obtained the following results:

Automatically extracted correctly-formed phrases from text: 171

existing bigram and 5-gram correctly formed phrases in text: 93

existing trigram and quadrigrams correctly formed phrases in text: 189

The recall value of the existing correctly formed trigram and quadrigrams is 90% (171 ÷ 189), and the overall recall is 60% (171 ÷ 282); Note that the total number of existing phrases in the above-mentioned senetnces is 282 (189 + 93). The reason for not very high performance is due to ignoring bigrams and 5-grams. Moreover, 10% of trigrams and quadrigrams could not be extracted by the proposed patterns.

There exist some works in the literature which attempt to extract key-phrases [29][30] from text, but those phrases are different from the ones extracted in the current work because any potential phrase (not only key phrases) are extracted in the current work.

In order to extract the above-mentioned phrases, an NLP tool for Turkish named ITU parser [31] is exploited to assign POS tags to the words in a sentence. Other NLP techniques such as lemmatization or normalization were not used.

Note that the collocated expressions are not necessarily compositional. As defined by Manning and Schütze [32], an expression is compositional if its overall meaning can be estimated based on the meaning of its parts. For example, the meaning of non-compositional phrase, "göz boyamak" (to deceive), cannot be estimated according to its words (literally means coloring the eyes); so, knowing that this phrase has negative polarity and catching it in the text helps estimate the polarity of the text including the phrase.

B. Basic Features for Phrase Classification

The list of features for phrase classification is provided in Table II, and explained below.

- N-grams: This method computes the co-occurrence probability of terms (words) with each other in a phrase. The goal is to distinguish correctly formed phrases from incorrectly formed ones. If the co-occurrence probability of included terms in a phrase is high, most probably they constitute a correctly formed phrase. As mentioned in [32], N-gram language model can be computed by probabilities given in Eq. (1).

$$log(P(t_i t_j t_k)) = log(P(t_i)) + log(P(t_j \mid t_i)) + log(P(t_k \mid t_i t_j)) \quad (1)$$

  $P(t_i)$ is the probability of seeing the term t_i in a phrase, $P(t_j \mid t_i)$and $P(t_k \mid t_i t_j)$ are respectively conditional probabilities of seeing t_j and t_k after seeing the given terms t_i and $t_i t_j$ in a phrase, and $P(t_i, t_j, t_k)$ is the probability of having correctly formed phrase with three terms: t_i, t_j, and t_k. For example, in the phrase "daha fazla olmalıydı" (it should be more (better)), extracted by the pattern [Adv Adv Verb], *log(P(daha))*, *log(P(daha|fazla))*, and *log(P(olmalıydı|daha fazla))* are computed. A similar equation could be written for quadruples.

- Hit number in a search engine: In this feature, each phrase is searched in Google search engine to capture its hit number. The higher the number of hits for a phrase, the higher the probability of correct formation.
- Document frequency: This feature counts the number of times each phrase appears among 270,000 Turkish sentences (unlabelled) as Turkish movie reviews.

After training a classifier by using the above mentioned features, all phrases are classified as "correctly formed" and "incorrectly formed". By the help of this classification, incorrectly formed phrases are removed from the list. This classifier has been trained by 1,000 phrases manually labeled as "correctly formed" and "incorrectly formed". The labeling task has been done by two (plus one) native Turkish speakers. The agreement of two labelers is 85.4%, and the third labeler helped

in labeling 14.6% of phrases which were not agreed by two labelers. The input of this classification task is a set of 5213 phrases extracted by the patterns of Table I and the output is a set of 4950 correctly formed phrases. A correctly classified sample is "üstüne yok doğrusu" (Actually there is no higher level upon it) and an incorrectly formed phrase which was misclassified as correctly formed is "bir film günün en ..." (the most ... of a movie day). Note that an incorrectly formed phrase very unlikely appears in a Turkish sentence. Also in some cases, not all words of a phrase are extracted by the proposed patterns; extracting only some (not all) words of a correctly formed phrase makes it incorrectly formed. The classification method used in this work is Logistic classifier which is used for its high generalization accuracy; the classification tool is WEKA, which is a known java-based machine learning tool, and the evaluation method is 5-fold cross-validation. In this evaluation method, the training set is divided into five equal parts, the first four parts (80%) are used as training set and the remaining 20% of data are supposed as test set. This task is repeated for five times for different 80/20 percent of training data.

TABLE II. Features Extracted for Classifying Phrases As Positive, Negative, or Neutral

Phrase Extraction	N-grams Hit number in Google Document frequency
Polarity Classification	Appearing in Pos/Neg sentences Pos/neg word count

C. Features for Phrase Polarity Classification

The classification features for phrase extraction and polarity classification are listed in Table II. First set of features have been used for phrase extraction (explained in Section III.B) and the rest of features have been used for polarity classification of phrases which are explained below.

- Appearing in Positive/Negative sentences: This feature counts the number of times a phrase appears in 2,700 positive and negative sentences--as a subset of movie reviews. The details of this subset are given in Section IV.A.
- Positive/negative word count: This feature captures the number of positive and negative terms appeared in a phrase. Two Turkish polarity lexicons are used for this purpose: Polar Word list and SentiTurkNet. In polar word list, words are already separated as positive and negative; In SentiTurkNet, similar to SentiWordNet, three polarity scores are assigned to each Turkish synset. A Turkish word is assumed as positive (or negative) if the average positivity (or negativity) score of its synsets is greater than their average negativity (or positivity) score. This feature is assumed as a baseline for phrase lexicon generation as it simply counts the number of positive and negative terms in a phrase.

D. Polarity Classification of Phrases

After classifying each phrase as *correctly formed* or *incorrectly formed*, the *correctly formed* phrases are classified as positive, negative, or neutral. For this purpose, two classification tasks (listed below) are carried out by using features listed in Table II. Similar to the first classification task, the classifier, evaluation method, and classification tool are respectively logistic regression, 5-fold cross validation, and WEKA.

- Classifying phrases as subjective and objective (neutral): In this classification, the output list of the phrase extraction phase (correctly formed phrases) is classified as objective and subjective; in other words, objective phrases are removed from the list. The input of this classification is a set of 4950 phrases and the output is a set of 2092 subjective phrases ignoring 2858 objective (neutral) ones. A correctly classified sample is "nasıl böyle saçma" (how silly like this) and an objective phrase which is incorrectly classified as subjective is "tabii romantik komedi" (of course a romantic comedy). The training set for this classification is a set of 800 correctly formed phrases which have been manually labelled as subjective and objective by two (plus one) native speakers, with 88% agreement on the labels of the two labelers, and getting help from the third labeller on 12% of labels which were not agreed at least by two labelers. The labels of remaining 4150 phrases (4950-800) are estimated by the trained classifier.
- Classifying subjective phrases as positive and negative: In this classification task, the output of previous step (subjective phrases) are classified as positive and negative. The input of this classification task is a set of 2092 phrases and the output is a set of 1591 positive and 501 negative phrases. The lower number of negative phrases is due to the lower number of negative reviews and sentences in movie reviews. The training set for this classification is a set of 500 correctly formed phrases which have been manually labelled as positive and negative by three native speakers of Turkish, with 83% agreement among two (plus one) labellers, getting help the third labeller on 17% of data which were not agreed by at least two labellers. The labels of remaining 1692 phrases (2092-500) were estimated by the trained classifier.

A correctly classified positive phrase is "tek işe yarar ..." (the only useful ...); a correctly classified negative phrase is "kesinlikle çok gereksiz bir..." (Absolutely a very unnecessary ...). A positive phrase that has been misclassified as negative is "izlediğim en iyi gerilim" (The best intensity movie that I have ever watched).

Note that instead of two binary classification (objective/subjective and positive/negative), one ternary (positive/negative/neutral) classification task has been also accomplished which is explained in Section IV.B.

IV. Experimental Setup

This section evaluates the proposed methodology by classification accuracy, extrinsic evaluation, and confusion matrix. Note that there is no Turkish polar phrase lexicon, therefore the generated list is new to Turkish. A subset of the generated lexicon is illustrated in Table III, and the complete and cleaned list can be provided for researchers via email. In this table, the column named "composition" shows the polarity of constitutive words in a phrase. A phrase may be composed of positive and objective (PosObj) words, negative and objective (NegObj) words, only objective words (Obj), or positive and negative words (PosNeg). Note that PosNeg phrases exist only in negative set; No positive phrase include both positive and negative words.

TABLE III. A Small Subset of Positive and Negative Phrases

Phrases	composition	tag
etkileyecek bir konu (an impressing subject)	PosObj	P
farklı bir eser (a different work)	Obj	P
olumsuz dersem yalan olur (I cannot say it is negative)	NegObj	P
büyük bir ayıp (a big shame)	NegObj	N
bir anlamı yok (does not have any meaning)	Obj	N
daha iyi olmalı (It should be better)	PosObj	N
iyi bir felaket! (a good disaster!)	PosNeg	N

We also investigated the distribution of positive, negative, and objective words in positive/negative phrases, which is illustrated in Fig. 3 and 4. In Fig. 3, x axis is the number (and percentage) of negative (or positive in Fig. 4) and y axis is the number (and percentage) of positive (or negative in Fig. 4) words in generated positive (or negative in Fig. 4) phrases. For example the number 33% in coordinate [0,0] of Fig. 3 means that 33% of positive phrases has zero positive and zero negative words.

As seen in Fig. 3, majority of positive phrases are composed of objective words, or objective plus positive words but they do not include negative words; however, positive words can be seen in negative phrases. In summary, it is usual to see positive (or negative) words in positive (or negative) phrases but the contribution of positive words in negative phrases (21%) is much more than the contribution of negative words in positive phrases (zero). Numbers upon each circle shows the percentage of phrases included in it.

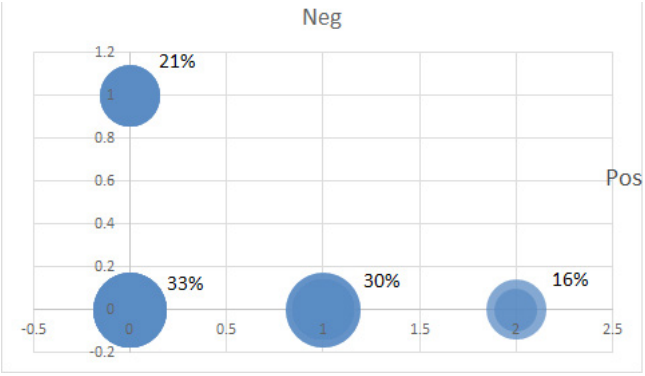

Fig. 3. Distribution of polar words in positive phrases. Numbers upon each circle show the percentage of phrases included in them.

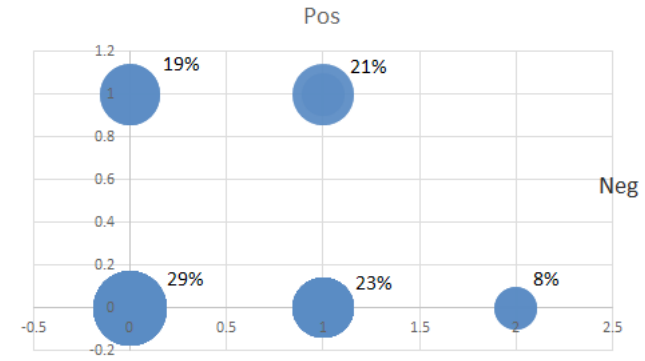

Fig. 4. Distribution of polar words in negative phrases. Numbers upon each circle show the percentage of phrases included in them.

A. Datasets

As mentioned in previous sections, the proposed approach has been applied on documents of movie domain in Turkish, which are more formal than some other data types such as tweets. Applying the suggested methodology on other kinds of textual data such as social media text would be challenging, as those data (e.g., tweets) are usually informal and noisy, including abbreviations and useless text.

Sentiment analysis is a domain-dependent task, therefore a given term may have different polarities in different domains; e.g., the term "big" is positive for *room size* in hotel domain but negative for *battery size* in camera domain. Although extracting polar terms (or phrases) from one domain and applying on another may have some drawbacks, however, in resource-lean languages such as Turkish we have to accept these weaknesses; moreover, some of the extracted phrases from movie domain are domain independent. In this work, two datasets have been used, first one for extracting the phrases and the second one for an extrinsic evaluation.

- Turkish movie reviews[4]. We have manually labelled 1,000 randomly chosen documents from this dataset as positive, negative, or neutral in our previous work [28]. We also labelled 2,700 sentences appearing in these documents as positive, negative, or neutral. Only the labels of sentences are used in this work. The distribution of [neutral, positive, and negative] sentences and documents are [50%, 30%, 20%] and [52%, 29%, 19%] respectively. The average length of each document and each sentence in this domain are respectively 23 and 9 words. The labeling task is accomplished by three (plus one) people and the agreement among at least two labellers is 81% for sentence level analysis. Again, the fourth labeller helped tag those sentences which were not agreed by at least two labellers. The already assigned rating scores to each movie review are not used in this work because we require labels in the sentence level but existing rating scores of movie reviews are available only at document level. We preferred manual labelling which is also more accurate than the rating scores. Each sentence or document is labelled as positive, negative, or neutral, if it conveys a positive, negative, or neutral polarity to the reader.
- Turkish restaurant reviews[5]. This dataset was used as training set in Semeval 2016-task 5 [33], which has been already labelled with three tags: positive, negative, and neutral. The aim of this task is to estimate the polarity label of each aspect appearing in a sentence. This dataset includes 239 documents and 1104 sentences, which has been used in the current work for evaluating the generated lexicons in sentence-level sentiment analysis. The average length of each document and each sentence in this domain are respectively 26 and 8 words.

B. Evaluation of Phrase Polarity Lexicon

In order to separate polar phrases from non-polar ones, one ternary (positive/negative/neutral) and three binary classification tasks (correctly/incorrectly formed, objective/subjective, and positive/negative) have been accomplished. The intuition behind this is that incorrectly formed phrases must be excluded from the extracted list, then the remaining list should be classified as positive, negative, or neutral. The classification accuracies for binary classification of phrases as correctly formed and incorrectly formed are listed in Table IV, and classification accuracies for binary and ternary classification of correctly formed phrases are listed in Table V. Moreover, confusion matrices for both binary (positive/negative) and ternary (positive/negative/neutral) classification of correctly formed phrases are provided in Tables VI to IX.

TABLE IV. Binary Classification of Turkish Phrases as Correctly Formed and Incorrectly Formed by Logistic Classifier Using 5-Fold Cross Validation On Training Data (%)

Feature name	correct/incorrect
N-grams	76.4
Hit number	70.45
Doc. freq.	72.20
All features	79.40

[4] These reviews are collected from www.beyazperde.com which are available in http://sentilab.sabanciuniv.edu/resources/

[5] http://metashare.ilsp.gr:8080/repository/browse/semeval-2016-absa-restaurant-reviews-turkish-train-data-subtask-2/ef952246940f11e5886b842b2b6a04d76a1959c4385a46bda776dd510ac3522e/

TABLE V. The Accuracy Of Binary And Ternary (Positive/Negative/Neutral) Classification of Turkish Phrases by Logistic Classifier Using 5-Fold Cross Validation on Training Data (%)

Feature name	ternary	subj/obj	pos/neg
pos/neg sentences	73.42	70.01	88.04
pos/neg words	71.02	68.22	85.16
Both features	74.43	72.90	91.31

TABLE VI. Confusion Matrix for Binary (Pos/Neg) Classification of Turkish Phrases with All Features

True	Estimated	
	positive	negative
Positive	0.93	0.07
Negative	0.18	0.82

TABLE VII. Confusion Matrix for Binary (Subjective/Objective) Classification of Turkish Phrases with All Features

True	Estimated	
	subjective	objective
Subjective	0.80	0.20
Objective	0.21	0.79

TABLE VIII. Confusion Matrix for Binary (Correctly/Incorrectly Formed) Classification of Turkish Phrases With All Features

True	Estimated	
	corr. formed	incorr. formed
corr. Formed	0.83	0.17
incorr. Formed	0.20	0.80

TABLE IX. Confusion Matrix for Ternary (Positive, Negative, and Neutral) Classification of Turkish Phrases with All Features

Feature name	Positive	Negative	Objective
Positive	0.79	0.05	0.16
negative	0.11	0.68	0.21
Objective	0.17	0.15	0.68

C. Extrinsic Evaluation

In order to evaluate the generated polarity lexicon, an extrinsic evaluation is carried out. The generated lexicon as well as other two Turkish lexicons, polar word set and SentiTurkNet, are used to estimate the polarity of Turkish restaurant reviews. This set includes 1104 Turkish sentences in restaurant domain. This dataset has been used as a benchmark in Semeval competition task 5 -Aspect based sentiment analysis. In this dataset, the goal is to estimate the polarity of aspects appearing in a sentence which have been tagged with three labels: positive, negative, and neutral.

The obtained accuracies with and without using the generated polarity lexicon are given in Table X. In this table, the abbreviations STN, PWS, and PL respectively stand for SentiTurkNet, Polar Word Set, and Phrase Lexicon. This sentiment analysis task simply searches for polar words in a sentence. No NLP technique except tokenization and word cleaning is employed in this system, as the goal is only to measure the usefulness of the generated lexicon. As seen in Table X, adding the phrase polarity lexicon increases the classification accuracy only by two percentage points. The reason (of low increment) can be the low number of idioms and multi-word polar phrases used in the sentences of restaurant domain. Moreover, catching a phrase in a sentence is not always straightforward. The appearance order of constituting words of a phrase in a sentence should be the same as its order in the phrase, so that the sentiment analysis system can find the phrase in the sentence. Note that phrases are extracted from movie domain but applied on restaurant domain. Extracting phrase from one domain does not necessarily makes them domain dependent. For example the phrase "nasıl böyle saçma" (how silly like this), can be used for any domain; however, there exist domain dependent phrases such as "iyi seyirler" (happy watching) which can be used only in movie domain.

TABLE X. The Accuracy of Binary (Positive/Negative) and Ternary (Positive/Negative/Neutral) Classification of Turkish Restaurant Reviews by Logistic Classifier Using 5-Fold Cross Validation on Training Data (%)

Lexicons used	Binary	Ternary
STN +PWS	73.02	67.23
STN+PWS+PL	75.17	69.22

V. Applying the Proposed Method on Other Languages

Since the grammar of natural languages is different from each other, in order to extract phrasal expressions from different languages, different patterns should be exploited. For example, in Turkish, verbs generally appear at the end of sentence, whereas in English, they usually appear in the beginning, after the subject. That is why in the suggested patterns for Turkish (Table I), verb is the last POS tag.

In this section, we examine how well the proposed methodology works on English. Below, necessary updates for applying the suggested methodology on English are explained.

Phrase Extraction: The suggested patterns in Table I should be adapted to English as done in Table XI. Note that the order of POS tags (especially verb) is changed. These patterns are used to extract candidate phrasal expressions from English movie reviews v2.0 [35]. As a result, 2588 raw phrases are extracted from the corpus.

Features and Classification: Classification process is the same as what was accomplished for Turkish. The first input of classification tasks is a set of 2588 raw phrases and the final output is a set of 295 negative and 534 positive phrases. In terms of features, two features, 'Hit number' and 'N-grams' are exactly the same as those used for Turkish, but remaining features use English resources. The feature 'document frequency' searches the generated phrases among randomly chosen 20000 sentences (ublabelled) from English movie reviews[6]. The feature 'appearing in pos/neg sentences' use 4200 sentences extracted from movie reviews, labelled as positive and negative. The feature, 'pos/neg words' benefit from three English polarity lexicons: SenticNet [17], SentiWordNet [18], and Liu's polarity lexicon [20]. In Liu's lexicon, positive words are separated from negative ones.

TABLE XI. Patterns of Table I Adapted to English. Numbers Inside Parentheses Show the Number of Phrases Extracted by Each Pattern

Triples	quadruples
verb adv adj (502) like very much	adv adj adj noun (286) very hard unsolvable problem
adv adj noun (372) very interesting effect	adv adv noun noun (226) very long time friend
verb noun adj (576) love you much	verb noun adv adv (191) put it somewhere else
verb adj noun (310) violate human rights	verb adv adv adj (125) support very very much

In SenticNet, we suppose a word as positive if its overall polarity score is greater than 0, or negative, otherwise; and in SentiWordNet, we suppose a word as positive if the average positivity polarity score

[6] http://www.cs.cornell.edu/people/pabo/movie-review-data/

of all synsets of the word is greater than its average negativity score, or negative, otherwise.

Evaluation: In this phase, we only provide classification accuracy for different classification tasks and omit other results. Table –XII and Table XIII provide these accuracies. As seen in Table XII, similar to Turkish, 'N-grams' is the most effective and 'Hit number' is the least effective feature; however, 'Hit number' feature in English has a little higher accuracy than in Turkish due to tremendous amount of English text in web compared to Turkish text in it. Moreover, overall accuracy in English is slightly greater than in Turkish.

TABLE XII. Accuracy of Binary Classification of English Phrases As Correctly Formed and Incorrectly Formed by Logistic Classifier Using 5-Fold Cross Validation on Training Data (%)

Feature name	correct/incorrect
N-grams	74.94
Hit number	72.05
Doc. freq.	73.32
All features	79.95

TABLE XIII. The Accuracy of Binary and Ternary (Positive/Negative/Neutral) Classification of English Phrases by Logistic Classifier Using 5-Fold Cross Validation On Training Data (%)

Feature name	ternary	subj/obj	pos/neg
pos/neg sentences	72.02	67.21	88.83
pos/neg words	69.22	66.92	85.66
Both features	72.33	70.80	92.01

In Table XIII, only binary classification of English phrases into positive and negative has slightly higher accuracy than the similar classification in Turkish, due to richer polarity lexicons and resources in English; other two classification tasks (subjective/objective and the ternary classification) have a few percentage points lower than the same tasks in Turkish.

VI. Discussion and Comparison

To the best of our knowledge, there is no work in Turkish to generate polar phrases, and other works have been applied on different datasets and languages (e.g., English). Therefore only similar works in English are reported below to provide a relatively fair comparison.

For generating polar phrases, Agrawal et al. [12] could achieve the accuracy of 70% in ternary (positive/negative/objective) classification and 84% in binary (positive/negative) classification of English phrases experimented on MPQA as the dataset. In [34], accuracies in neutral/polar classification range from 65% to 76% and 69% to 83% in polarity classifications for different datasets. In the current work, the best classification accuracies for ternary (positive/negative/objective), polar/objective, and positive/negative classification of Turkish phrases are respectively 74%, 73%, and 91%, whereas the same accuracies for English phrases are respectively 72%, 70%, and 92%. Due to different datasets used in the above-mentioned related work and the current one, the comparison may not be totally fair.

The most similar previous work to this one has been accomplished in [6]; the main difference between these two works is that the suggested approach in this paper for phrase extraction and annotation is semi-automatic but the annotation of phrases in [6] is manual (although phrase extraction is automatic and pattern-based).

According to the results reported in Section IV, the following conclusions can be extracted.

- The proposed approach for phrases, outperforms the baseline approach--counting the number of positive and negative terms in phrase--by 1 to 3 percentage points. This issue emphasizes the effect of non-compositional phrases in sentiment analysis, in which the polarity of the whole phrase cannot be estimated based on the polarity of its parts.
- The best classification accuracy in both Turkish and English phrases has been obtained in binary classification of phrases into positive and negative.
- In correctly/incorrectly formed classification of phrases, the N-gram feature obtained the highest accuracy. This finding approves the assumption that the higher the co-occurrence probability of a word-pair, the higher the probability of correct phrase formation by this pair.
- The highest per-class accuracies (confusion matrix values) belong to the positive class and lowest accuracies belong to the negative class. Generally positive expressions are more clearly expressed by people, compared to the negative expressions.
- Catching phrases and idioms in a sentence is not as easy as catching unigrams and bigrams in it as in some cases, phrases are separated by other words in the sentence.

VII. Conclusion and Future Work

In this work, a semi-automatic methodology is proposed to build phrase polarity lexicons. The proposed methodology consists of several methods such as word co-occurrence probability. Because the polarity of phrases cannot usually be estimated based on the polarity of its parts, covering phrases in sentiment analysis in a very challenging task. The generated lexicon is freely available for research community. Although the paper mostly focused on Turkish, the proposed methodology is language-independent and can be applied on other languages with small changes. The future work consists of adding polar idioms to existing polarity lexicons, and considering language issues such as negation in generated phrases.

References

[1] B. Liu, Sentiment Analysis and Opinion Mining, New York, NY: Cambridge University Press, June 2015.

[2] G. A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (11), 1995, pp. 39-41.

[3] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417-424.

[4] V. Hatzivassiloglou, K. R. McKeown, Predicting the semantic orientation of adjectives, in: ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997, pp. 174-181.

[5] J. Martineau, T. Finin, Delta tfdf: An improved feature space for sentiment analysis, in: Proceedings of the Third International ICWSM Conference, 2009, pp. 258-261.

[6] S. Kiritchenko, S. M. Mohammad, Happy accident: A sentiment composition lexicon for opposing polarity phrases, in: Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia, 2016.

[7] R. Dehkharghani, Y. Saygin, B. Yanikoglu, K. Oazer, Sentiturknet: a Turkish polarity lexicon for sentiment analysis, Language Resources and Evaluation. 50, 2016, pp. 1-19.

[8] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing – Vol. 10, 2002, pp. 79-86.

[9] A. Meena, T. Prabhakar, Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis, ECIR 2007: Advances in Information Retrieval, 2007, pp. 573-580.

[10] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: HLT '05 Proceedings of the conference

on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347-354.

[11] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, Computational Linguistics. 35 (3), 2009, pp. 399-433.

[12] A. Agarwal, F. Biadsy, K. R. Mckeown, Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, in: EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 24-32.

[13] J. Yi, T. Nasukawa, R. Bunescu, W. Niblack, Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, in: ICDM: Third IEEE International Conference on Data Mining. ICDM. 2003, pp. 427-434.

[14] T. Nasukawa, J. Yi, Sentiment analysis: Capturing favorability using natural language processing, in: K-CAP '03 Proceedings of the 2nd international conference on Knowledge capture, ACM, 2003, pp. 70-77.

[15] A. C.-R. Tsai, C.-E. Wu, R. T.-H. Tsai, J. Y.-j. Hsu, et al., Building a concept-level sentiment dictionary based on commonsense knowledge, IEEE Intelligent Systems 28 (2), 2013, pp. 22-30.

[16] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, S. Bandyopadhyay, Enhanced senticnet with affective labels for concept-based opinion mining, IEEE Intelligent Systems 28 (2), 2013, pp. 31-38.

[17] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and commonsense knowledge base for cognition-driven sentiment analysis, in: Twenty eighth AAAI conference on artificial intelligence, 2014, pp. 1515-1521.

[18] M. Hu, B. Liu, Mining and summarizing customer reviews, in: KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 168-177.

[19] S.-M. Kim, E. Hovy, Determining the sentiment of opinions, in: COLING '04 Proceedings of the 20th international conference on Computational Linguistics, 2004, Article no. 1367.

[20] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: LREC, Vol. 10, 2010, pp. 2200-2204.

[21] O. Bilgin, Ö. Çetinoğlu, K. Oazer, Building a wordnet for Turkish, Romanian Journal of Information Science and Technology 7 (1-2), 2004, pp. 163-172.

[22] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domainoriented sentiment analysis, in: EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 355-363.

[23] E. Yıldırım, F. S. Çetin, G. Eryiğit, T. Temel, The impact of NLP on Turkish sentiment analysis, Türkiye Blisim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi 7 (1) (Basılı 8), 2015.

[24] A. G. Vural, B. B. Cambazoğlu, P. Şenkul, Z. Ö. Tokgöz, A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish, E. Gelenbe, R. Lent (Eds.), in: ISCIS, Springer, 2012, pp. 437-445.

[25] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, Journal of the American Society for Information Science and Technology 63 (1), 2012, pp. 163-173.

[26] M. Kaya, G. Fidan, I. H. Toroslu, Sentiment analysis of Turkish political news, in: WI-IAT '12 Proceedings of The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01, 2012, pp. 174-180.

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

[28] R. Dehkharghani, B. Yanikoglu, Y. Saygin, , K. Oazer, Sentiment analysis in Turkish at different granularity levels, Natural Language Engineering 23 (4), 2017, pp. 535-559.

[29] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, C. G. Nevill-Manning, Kea: Practical automatic keyphrase extraction, in: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, 2005, pp. 129-152.

[30] Y. Zhang, N. Zincir-Heywood, E. Milios, Narrative text classi_cation for automatic key phrase extraction in web document corpora, in: Proceedings of the 7th annual ACM international workshop on Web information and data management, ACM, 2005, pp. 51-58.

[31] G. Eryiğit, ITU Turkish NLP web service, in: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics, Sweden, 2014, pp. 1-4.

[32] C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT Press Cambridge, MA, USA, 1999.

[33] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, Orphee declercq, v_eronique hoste, marianna apidianaki, xavier tannier, Natalia loukachevitch, evgeny kotelnikov, nuria bel, salud mar_a jim_enez-zafra, and gülşen eryiğit. semeval-2016 task 5: Aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval, Vol. 16, 2016, pp. 19-30.

[34] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis, Computational linguistics 35 (3), 2009, pp. 399-433.

[35] B. Pang and L. Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, pp. 271.

Rahim Dehkharghani

Dr Rahim Dehkharghani received his PhD in Computer Science and Engineering, Artificial Intelligence branch, from Sabanci Universty, Istanbul, in 2015 and his MSc degree in Computer Engineering, Software branch from Shahid Beheshti University, Tehran, in 2007. His main research area is Natural Language Processing and Data Mining. He specifically works on sentiment analysis of textual data. He is a faculty member of Computer Engineering group at University of Bonab (Bonab, Iran) since 2016.