DETECTING SUBJECTIVITY IN THE NEWS TEXTS IN TURKISH LANGUAGE

by

Dicle Öztürk

B.S., Computer Engineering, İstanbul Technical University, 2008

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2014

ACKNOWLEDGEMENTS

This study has made me involve in the process where doing research for itself as a student and selling the labour in itself as a teaching assistant interlace so surprisingly as if they were two completely irrelevant occupations. This is what the majority of the researchers are going through almost all around the world. What I mean in other words is that research is getting to be a luxury, year by year. I wish we hadn't witnessed or just encountered this deterioration but instead endeavoured to change our work conditions. I would like to deliver a self-criticism for not having strived enough for our conditions and rights.

Whatever it took, and maybe it was all for this pleasure, I had also the great chance of getting deeper on a subject, creating a system from scratch and coming up with new problems while solving others. For helping me pursue such an occupation that nevertheless diminishes the amount of alienation in life, I would like to thank my beautiful family, all of my dear friends and dear comrades and my professors with all my respect and loving.

I dedicate all my work to Ceylan Önkol.

ABSTRACT

DETECTING SUBJECTIVITY IN THE NEWS TEXTS IN TURKISH LANGUAGE

Subjectivity and sentiment analysis research has gained increasing attention in the recent years like many language technologies. Its aim is to investigate and to develop techniques to recognize subjectivity or sentiment in human-generated content such as text, speech or image. While subjectivity and sentiment detection tasks are necessarily related to each other, subjectivity detection is relatively understudied and needs more attention, being a challenging problem even for humans. For capturing subjectivity clues in the text, various linguistic properties are made use of and for predicting the subjectivity of an unknown piece of text, machine learning methods are applied. In this respect, the subjectivity detection problem can be reduced to a text classification problem. A set of texts evaluated for some predefined clues of subjectivity, are input to a learning module, which will predict if a given unknown piece of text is subjective or objective. In this work, we study subjectivity detection in news items using machine learning methods and develop a framework that runs at the document-level. We assume that the descriptive features of expressions is a good candidate to capture the subjective tone in texts and based on this premise, propose a novel feature set for subjectivity classification. We implement a supervised scheme and extensively evaluate it on a dataset which we have collected and annotated. Our findings present new directions and useful contributions to the subjectivity detection literature. We introduce the first subjectivity detection system in Turkish language, present our new database with annotations and report high accuracy in subjectivity detection.

ÖZET

TÜRKÇE HABER METİNLERİNDE TARAFLILIK TESPİTİ

Taraflılık ve olumluluk analizi, son yıllarda oldukça ilgi çeken bir araştırma alanı haline geldi. Bu alanda, temel olarak, metin, konuşma veya resim gibi içeriklerde taraflılık veya olumluluk gibi özellikler olup olmadığını bulmayı sağlayan yöntemler geliştirilir ve araştırılır. Taraflılık ve olumluluk analizi alanları, adlarının da çağrıştırdığı gibi, birbirleriyle oldukça ilgilidir fakat taraflılık analizi görece daha az ilgi görmüş bir alandır ve insanlar içinbile zor bir konu olmasından dolayı daha fazla çalışılmaya muhtaçtır. Taraflılık tespitini ilk alt probleme ayırabiliriz; birincisi, taraflılık özelliklerini çıkarmak ve ikinci olarak, verilen yeni bir metnin taraflılığını tahmin etmek. Birinci problem için, dilbilimsel özellikler başlıca başvuru kaynaklarındandır. Taraflılığın tahmininde ise çoğunlukla yapay öğrenme yöntemleri kullanılır. Bu açıdan, taraflılık tespiti problemi, bir çeşit metin sınıflandırma problemine indirgenebilir. Biz bu çalışmada, yapay öğrenme yöntemlerini kullanarak, haber metinlerinde taraflılık tespiti problemi konusunda çalıştık ve doküman seviyesinde çalışan bir uygulama geliştirdik. Metinlerdeki betimleyici öğelerin taraflı tonu yakalamada iyi bir özellik olabileceği önkabulü altında, taraflılık sınıflandırması için yeni bir öznitelik kümesi tanımladık. Denetimli öğrenme algoritmaları kullanarak, yöntemimizi kendi topladığımız ve etiketlediğimiz bir veri seti üzerinde test edip değerlendirdik. Yöntemi-mizin ve deney bulgularımızın, taraflılık tespiti alanına katkı sunacak nitelikte kullanışlı olduğunu gördük; deneylerdeki başarım değerlerinin de düşük olmadığını gözlemledik. Sonuç olarak, bu çalışma ile Türkçede yapılmış ilk taraflılık sınıflandırması sistemini, etiketlenmiş yeni bir veri seti ile beraber sunuyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

b	Intercept term of a hyperplane
c_i	The ith class
C	The regularization or the penalty parameter to prevent over-fitting
k	The length of the vector $\mathbf{x}_i$
K	The set of prediction classes
$\mathbf{w}$	The weight vector perpendicular to a hyperplane
$x_{i,j}$	Value of the jth feature of the ith data point
$\mathbf{x}_i$	The ith data point as a vector
y_i	Class value of the ith data point
α	Krippendorff's inter annotator agreement rate measure
κ	Cohen's inter annotator agreement rate measure
λ_i	The ith Lagrange multiplier
ξ_i	The ith slack variable
τ	The smoothing parameter in Multinomial Naive Bayes

LIST OF ACRONYMS/ABBREVIATIONS

ANEW	Affective Norms for English Words
BoW	Bag of Words
ISEAR	International Survey on Emotion Antecedents and Reactions
LIWC	Linguistic Inquiry and Word Count
MPQA	Multi-Perspective Question Answering
POS	Part of Speech
RBF	Radial Basis Function
SO	Strongly Objective
SS	Strongly Subjective
SVM	Support Vector Machine
WO	Weakly Objective
WS	Weakly Subjective

1. INTRODUCTION

Being a part of the language understanding process, identifying attitudinal and subjective content in an automated (computerized) way is a hard but seemingly not an impossible task, as much as most of the computational intelligence applications. To investigate the possibility for the computers to predict subjectivity, we first need to clarify what we understand from subjectivity. Let us consider the case where someone says "The weather is nice". This is a subjective sentence because the judgement of the weather differs from person to person. However, when someone says "The weather is rainy", we can easily say that to the universal common sense, this sentence can be regarded as objective because it only reports the current situation and anyone with the ability of seeing can validate it. In that sense, it appears that objective measuring mechanisms are an important factor in the evaluation of subjectivity and humans can easily distinguish those mechanisms.

Automated recognition of sentiment and subjectivity has gained importance over time, due mostly to the tremendous accumulation of all kinds of content on the Web [1]. Various approaches and techniques have come out in the field, extending the capabilities of the computers to predict the attitude of a piece of text. Sentiment analysis is a branch of text mining, which detects and analyses sentimental, emotional or polar (positive, negative, neutral) clues in texts, using statistical or rule-based techniques, supported by linguistic information (See [1] for a comprehensive survey.). Subjectivity detection in texts differs from sentiment analysis, in that it investigates the methods of distinguishing objective or factual voice from the subjective tone in texts [2–7]. The medium of information ranges from online product comments to movie reviews, from news stories to blog entries. The aim is to automatically obtain an agent's attitude or sentiment toward some other party by mining through the target texts.

In this work, we have built a supervised subjectivity learning framework, where subjectivity features of texts are extracted from their lexical and part of speech (POS) tag attributes. We propose a novel feature extraction method for subjectivity clues

and show that it produces promising results. The ground truth of the system is based on a set of labelled news texts, which is obtained with an annotation scheme that we have developed. To better understand the dimensions of a subjectivity learning system, we also performed several experiments comparing different features and classifier models. We expect this study to offer a new approach to detecting subjectivity as well as being useful in the related fields such as information extraction and word-sense disambiguation as a by-product.

The outline of this work is as follows: In Chapter 2 (Learning Subjective Language), we summarize related work on automated subjectivity and sentiment detection and introduce the theoretical basis of our approach, in addition to listing our primary research questions. Chapter 3 (Dataset) describes the Turkish newspaper corpus we have collected and annotated, as well as its subsets. Chapter 4 (Methodology) explains our feature extraction and classification approach. Chapter 5 (Experiments) describes the experimental setting and discusses the results of our experiments. Finally, in Chapter 6 (Conclusion & Future Work), we complete this work with a general overview and comments on future work.

2. LEARNING SUBJECTIVE LANGUAGE

2.1. Subjectivity, Sentiment, Opinion, Polarity

Natural language processing literature prominently employs the subjectivity definition that Wiebe and Rapaport introduced to the literature in 1988 [8] and Wiebe further detailed in her work in 1994 [9]. In their 1988 work [8], which is actually a study on the computational analysis of narratives, Wiebe and Rapaport convey a definition of subjectivity that was borrowed from Banfield's work titled "Unspeakable Sentences: Narration and Representation" [10]. Banfield says that a narrative sentence is objective if it transfers the story as it is, without putting the commentary of the character. And it is a subjective sentence if the thoughts, perceptions, or more compactly, perspective of the character are conveyed along with the story [10]. In 1994 Wiebe gives a more elaborate description of subjectivity, connecting Banfield's approach to the concept of private states [9]. She claims that a sentence containing private states is subjective. There, following the definition given by Quirk *et al.* [11], Wiebe reports that private states are the states of mind that cannot be observed by others. Some examples are listed such as 'knowing', 'believing' as intellectual private states, 'hating' as an emotive private state or 'hearing' as perceptual private state. All in all, Wiebe induces that when Banfield categorizes a sentence as subjective, it is actually subsumed by the concept of private states, namely private state descriptions that depict the character's psychological condition or "point of view".

As for the definitions of sentiment, polarity and opinion; although their dictionary meanings are not quite overlapping, we see that these three terms denote the same concept in the literature. Roughly, sentiment, opinion or polarity of a text is the value that a source loads on a target. For example, the sentence "I liked this book" carries a positive sentiment held by "me" towards "this book"; in other words, polarity of the sentence is positive or the opinion conveyed in the sentence is positive towards "this book", where the opinion belongs to me. In that regard, we can say that sentiment is a more specific feature than subjectivity.

To our knowledge, although there is not a sharp consensus over the issue, it is mostly assumed that content that is positively or negatively sentimental is subjective. In that sense, subjectivity detection and sentiment analysis / opinion mining fields are highly interrelated. There are some exceptions claiming that an objective textual unit can well have positive or negative polarity values [7] or preferring to have subjectivity and sentiment analyses separate tracks [2–4]. Pang and Lee present a framework that starts with subjectivity classification and feeds its results to sentiment (polarity) detection, reporting to have improved results this way [5].

2.2. Related Work

We will take both sentiment analysis and subjectivity detection literature into account in our short survey to give a sense of subjectivity detection research related to our problematic, as well as to present the categorical differences that we have noticed.

2.2.1. The Media of Textual Data

The first distinction pertains to the textual resource being analysed. The sources can be informal online chat records [6, 12], discussion portals [13, 14], news stories [15–20], social network communication [4,21], blog entries [22], product reviews [23,24] and movie reviews [3, 5, 6, 23–26]. One technique to analyse polarity or subjectivity may result in differing performance scores over different sources of text [6]. Blitzer *et al.* study domain adaptability of sentiment analysis systems, as well as developing a measure of domain similarity, which is useful in adapting a classifier from one data domain to another [27]. Another study investigating the cross-domain performance of sentiment analysis systems can be found in [28] where the developed system is said to have given promising results in extracting a polarity lexicon from unlaballed texts of a domain, using an existing lexicon from another domain's texts.

Studies in this field generally handle English texts. The lack of resources in non-English languages can be attributed to a number reasons, which we will not discuss here. A survey on sentiment and subjectivity analysis in Arabic texts can be found

in [29]. A study on the sentiment classification of Czech texts is given in [30]. We present studies dealing with Turkish texts in Section 2.2.5. Some cross-language studies are also available, discussing the applicability / speciality of sentiment / subjectivity analysis systems over different languages [26,31–35], using machine translation modules or probabilistic models of languages.

2.2.2. The Units of Text

The studies differ in terms of the unit of text that the methods are applied on. Polarity or subjectivity values of texts can be evaluated at the document-level [4, 14, 15, 24, 36–39], at the sentence-level [6, 7] or at the phrase-level [30, 40, 41]. A comparative picture of document-level and sentence-level analysis can be found in [13] and [15], both indicating that document-level prediction outperforms the latter. Pang and Lee [5] report an increase in accuracy when a sentence-level subjectivity analysis module in interposed into a target document-level sentiment detector.

2.2.3. Text Categorization Approaches

The studies differ in terms of their classification approach; they use either statistical or rule-based methods. Purely unsupervised learning methods generating Bayesian models of texts to predict sentiment are reported to give promising results [23, 24]. There are weakly-supervised approaches as well, which employ human intervention or bootstrapping [6, 7, 42].

In the supervised learning settings, we see that Support Vector Machine (SVM) and naive Bayes algorithms are reported to give very good results [5, 13, 15, 43–45]. Maas *et al.* present a supervised sentiment classification system that contains an intermediary unsupervised module as a helper for feature extraction [3].

The performance of an overall detection model might be influenced by the chosen classifier. For example, in [13], Lin *et al.* build two Naive Bayes models and compare their performance with SVM. They show that their naive Bayes models outperform

SVM on classifying documents by their ideological perspectives (pro-Palestenian or pro-Israeli) conveyed. Likewise, in [5], Pang and Lee experiment with a graph-based polarity detection algorithm with SVM and naive Bayes learners, reporting that in many settings, naive Bayes results in better performance than SVM. Indeed, it is also reported in [15, 44, 45] that naive Bayes algorithms fit well on the tasks of sentiment and subjectivity detection.

Lehrman *et al.* apply four different supervised classifiers, namely, naive Bayes, Max Entropy, Decision Tree and Max Vote, respectively in a study on detecting distressed affect states in short forum texts [14]. They obtain accuracy around 80% with all the four algorithms when they classify user forum posts as distressed or non-distressed; however, when they apply classification on four labels, taking the intensity of the binary labels into account, performance drops to around 60% and it is below 40% with the Max Entropy classifier. Naturally, a four-class classification problem is harder than a two-class problem. The authors make extensive use of linguistic features such as part of speech variables, unigrams, pronouns as well as using a polarity and an emotion lexicon, reporting the effect of feature sets as an another experiment, as well.

There are some studies that rely purely on dictionary-based methods [4,12,17,25] where the categorization is based on sentiment values of textual units, which measure their lexical properties using sentiment-related lexicons or WordNet [46]. [22] give a comparative account of machine learning algorithms and lexicon-based models for sentiment classification in blogs.

Some studies employ more exploratory ways to discover subjective voice in text using graph-based methods and visualization rather than experimentally detecting it [20]. For topical and discursive explorations of the news, in [47–49], very large visualization systems for news texts are developed and presented, which can inspire subjectivity analysis as well.

2.2.4. Textual and Linguistic Features

All learning algorithms require datasets be converted feature spaces where the classification will be performed. The choice of feature sets is an important part of the overall categorization process. In most of the studies that we have mentioned above in this section, a number of feature types are used and, in some, reported comparatively. Term weighting, which maps words onto the vector space, [3,19,21,42]; term weighting with polarity or affect lexicons [4, 16, 24]; N-Gram models of the content [15, 18, 19, 26, 43]; word classes (part of speech tags) [14, 15, 23, 50]; syntactic dependencies [4, 32] provide various contributions for feature extraction mechanisms. As a rather rare approach, in [18], the authors use the stylistic properties of words to predict subjectivity without using any lexicon or learning algorithm.

Some of the widely-known sentiment / affect / emotion dictionaries, applicable in English language, can be listed as follows; Subjectivity Lexicon of the framework Multi-Perspective Question Answering (MPQA) [40], MPQA Subjectivity Sense Annotations [51], LIWC (Linguistic Inquiry and Word Count) [52], General Inquirer [53], ANEW (Affective Norms for English Words) [54], SentiWordNet 3.0 [55], FrameNet [56]. Some discussions on the usefulness of affect dictionaries and the effect of words in emotive meaning can be found in Osherenko and André [57] and in Pennebaker *et al.* [58].

2.2.5. Sentiment Analysis of Texts in Turkish Language

To our knowledge, the first sentiment analysis study on Turkish texts is a 2009 master thesis entitled "Sentiment Analysis in Turkish" [59]. In this work, Eroğul classifies Turkish movie reviews as positive or negative and reports a comparative treatment of some feature sets. Words and the roots of the words are used as features separately at both sentence and document level. It is said that using roots results in slightly better performance. This study includes measuring the effect of part of speech (POS) tags where among taking single POS tags, verbs are reported to be the tag giving the highest accuracy. The author indicates that adjectives together with nouns give the best results for combined tags. As for the language modelling features, Eroğul

has studied N-gram models and shown that performance is inversely proportional to N, such that unigrams give the best results and trigrams decrease the accuracy remarkably. Overall, Eroğul reports an accuracy value of 85% using SVM classifier in their binary sentiment classification system, with more than 10000 movie review documents.

In another unpublished thesis, which is entitled "Emotion Analysis of Turkish Texts by using Machine Learning Methods" [50] in 2012, Boynukalın develops a supervised system to classify texts in Turkish language by four emotion dimensions, namely, joy, sadness, fear and anger, respectively. Two types of datasets are combined to test the proposed approach; one is a translation of some selected texts in ISEAR (International Survey on Emotion Antecedents and Reactions) [60], which contains survey answers describing the respondent's reaction to seven basic emotions, and the other dataset contains manually emotion-annotated sentences from 25 fairy tales in Turkish. The author employs unigram, bigram and trigram values of words in combination and using different weighting schemes. Additionally, both SVM and Naive Bayes algorithms are applied. While reporting comparative accounts of feature sets, measuring mechanisms and algorithms, the author states that the system reached an average accuracy around 69-81% on four-class classification.

In their 2013 work, Demirtaş and Pechenizkiy study supervised polarity detection in both English and Turkish texts using a cross-lingual approach [26]. Their dataset is made up of movie and product reviews in English and in Turkish separately. They investigate as to whether expanding the monolingual training set with texts automatically translated from other language increases classification performance or not. The answer in their setting is that the performance does not increase and the authors attribute this effect to "*cultural or other differences*" in the same domain of texts. Their system learn texts on the features of unigrams and bigrams of words along their binary polarity values, using linear SVM, Naive Bayes and Maximum Entropy classifiers; the system reaches up to an accuracy value of 86%.

Vural *et al.* present a crawler that proceeds through selecting texts from the Web by their polarity values [61]. The system is actually introduced in Vural's 2013 PhD

dissertation [62]. The sentiment module of the thesis is given in [63] where Vural *et al.* present an unsupervised lexicon-based sentiment detector with experimental results on Turkish movie review texts. In this work, firstly, the sentiment scores of the sentences in the corpus is obtained using a Turkish-customized version of the sentiment analysis tool SentiStrength [64]. The authors implement three different decision measures to output the sentiment value of a text. The best accuracy value that their system with two sentiment classes attains is reported to be 76%.

In [19], Kaya *et al.* classify newspaper column articles on politics by their sentiment value. Their corpus contains 200 positive and 200 negative articles, annotated by three persons. For sentiment learning, stemmed unigram, surface form unigram, unigrams with adjectives, unigrams with effective words which they themselves collected and lastly bigrams are used as features, while frequency and binary presence measures are employed separately. The algorithms applied are SVM, Naive Bayes, Maximum Entropy and N-Gram character language models. They show that stemming decreases performance and unigram model results better than bigrams. Among the algorithms, Naive Bayes gives the least accuracy as 72%. For the other three algorithms, the best accuracy is around 76%. Another result is that binary presence weighting scheme for feature vectors remarkably outperforms frequency weighting. In the paper, the authors say that they did not exclude stop words from the texts in feature calculations on the grounds that most of these stop words carry sentiment value. However, it is not validated as to whether with or without stopwords feature vectors give better results or not.

Another study on sentiment analysis in Turkish in 2013 proposes a new method that seeks to diminish the size of the labelled training set size while preserving the learning performance [21]. In this study, Çetin and Amasyalı use a supervised term weighting scheme to vectorize their corpus containing polarity-labelled Twitter data. Their initial result is that that their scheme far outperforms the ordinary unsupervised weighting methods. As for learning, they try a method called "active learning" that proceeds by selectively learning the training instances so that the need for labelled items can be decreased. Thus, experimenting with different selection mechanisms and

different classifiers, they show that active learning approach reaches the best result with an accuracy value of 64%, while requiring half of the labelled items compared to other mechanisms at the same performance level.

2.2.6. Subjectivity Analysis at the Document-level

Yu and Hatzivassiloglou present an opinion/fact classification experiment at the document-level where they use naive Bayes algorithm as classifier and the dataset is the news articles from the opinionated editorial articles and from the factual, reported (non-opinionated) news texts. They report this experiment as an additional task to sentence-level polarity classification. As for features for the document-level classification, they use the bag of words representation of words in texts, without stemming and stopword elimination. They report 97% for F-measure as the result of their experiment. We would like to note that this is a rather different task from subjectivity analysis in that the texts in the opponent classes are written highly different style and thus, they are more easily distinguishable than the texts from similar domains (such as the texts from the local and political news categories) as in our case.

Wiebe *et al.* give a vigorous study of subjectivity elements in texts, presenting some learning experiments that investigate the effect of 'potential subjective elements' on the precision of subjectivity identification [39]. They additionally report a supervised document-level opinion prediction; i.e., the task is to predict if a given text is opinionated or not (like neutral or not). The dataset is compiled out of news articles such that the texts from the categories 'editorial', 'letters to editor', 'arts review' and 'viewpoints' are automatically in the class 'opinion pieces' and the remaining texts are said to be in the class 'nonopinion pieces'. Thus, no manual annotation is needed. The only feature of text is the number of 'potential subjective elements' in the text divided by the number of words of the text. 'The potential subjective elements' are said to be the adjectival and verbal subjectivity clues they learned from the texts with various experiments. They use k-nearest neighbour classifier and report that the accuracy attained is 0.939 while the baseline (majority classification) is 0.915.

In [37], the authors apply review identification to a set of texts containing two classes each having 2000 documents; the class review with movie reviews, and the class non-review with plot summaries and ads. The features are the binary presence values of all the words and the classifier that they used is SVM. They report to have achieved 99.8% accuracy. They apply this model on book reviews as well; although the dataset size is halved, the resultant accuracy is reported to be 96.8%. We can say that review and non-review distinction is not as complex as subjectivity distinction. Reviews quantitatively and obviously differ from non-review texts by their word usage such as containing specific description words or phrases. This is not the case for our dataset domain and classes. Subjectivity requires qualitative distinction of the words. In fact, using binary term weighting as in this study, we have also tried classifying subjectivity annotated Turkish news texts in our dataset as subjective and objective and have obtained an accuracy of 56%, which is too low compared to this study.

In [38], Toprak and Gurevych report the results of their work developed within a challenge task on document-level subjectivity detection. They make use of binary weighted unigrams and bigrams, tf*idf (term frequency and inverse document frequency multiplication) weights of words as well as word window representations to take the context of a word into account. Additionally, they use subjectivity lexicons as features, calculating both binary presence and total occurrence values of them in text. Lastly, they employ POS tag features such as taking the count of pronouns, adjectives and adverbs in the text, windows of words with binary POS information and modal verb presence in the sentences. They have two sets of size larger than 6000 documents, one in French and the other one in English, both compiled from news articles and automatically annotated such that a text is subjective if it is from the category editorial and objective if it is a news texts from the categories such as local, politics or economy. They use SVM for classification and conduct several document-level prediction experiments investigating the performance of different feature types and measures. The best accuracy they attained is 0.855, which is the result of the experiment where the only feature used is the tf*idf weights of a word and its left and right neighbour words.

Scholz *et al.* develop an opinion corpus in German, which contains press releases

of political parties, manually annotated for sentiment. They apply subjectivity classification, as well as predicting sentiment [36]. Their textual unit is not the whole document of but statements which are made up of more than one sentence. They make use of four features for measuring subjectivity. Using a seed set of annotated statements, they calculate the chi-square, pointwise mutual information and association rule mining values of each word to evaluate their subjectivity inclination (or sentiment). They additionally use the tf*idf weight of each word as features. These four measures are used separately for each experiment. Applying naive Bayes algorithm for classification, they find that chi-square is the best performing measure, giving 71.74% accuracy. Not having a seed set of documents, for comparison purposes, we applied only bag of words style tf*idf weighting as features and classified the subjectivity-annotated news texts in our dataset using both multinomial naive Bayes cross-validated, the resultant accuracy is 69.06% ((±2.86)). The authors report to the accuracy to be 67.99% when only tf*idf weights as features is used, which is slightly higher than our result. Omitting the effect of the language-specific tools like stemmer, we can say that our annotation quality is comparable to this study where the texts are labelled by professionals.

2.3. Research Questions

2.3.1. What makes a piece of text subjective?

We firstly highlight that in the literature subjectivity is generally assumed to be a property ascribed to sentences or phrases only [65]. In fact, there are relatively few works studying subjectivity at the document-level, which we present in Section 2.2.6. In these studies, subjectivity detection is specified as an additional task to the sentiment analysis of the texts. We propose that, just as sentiment, subjectivity might also be analysed at the document-level and that it can be a sub-task of text classification. As a premise of this claim, we handle subjectivity as a property of texts concerning the biased tone contained, rather than a mental standardization of the verbal expressions. Embracing a more discourse-analytic perspective, we regard text as the unit of discourse and subjectivity as an attribute of this unit [66]. All in all, it is the texts that we classify as subjective or objective, not sentences or phrases, as

opposed to the majority of the studies on subjectivity given in Section 2.2.

In an attempt to capture new clues for subjectivity, we will try to form a more concrete base than the phenomenological stance of *states of mind - private states* approach to subjectivity, considering the specialties of our medium of data, namely the news text, as well as the less-advantaged situation of Turkish language resources and our philosophical objection to private state descriptions, which we will briefly rationalize here.

With regard to our philosophical concerns over the private-state descriptions, our basis is formed over the thoughts of Wittgenstein [67]. According to Wittgenstein, because we misunderstand language, our thinking that sensation terms refer to objects inside us is wrong. He remarks that in language there are words for sensations; however those terms do not describe them, only express them. Sensation is only knowable to oneself, thus it is private. However, as soon as we name a sensation, it becomes public and publicly justifiable through behaviour; the behaviour which stems from the sensation and which informs others about the sensation. Thus, we do not accept the property 'being not observable by others', which is attributed to the private states, because what is not observable about private states is their quality, not their existence. They are observable by the behaviour of the holder; i.e., they are known only when they are observed.

It is our intuition that a more pragmatic approach will bring realistic and computationally interesting dimensions. We propose that subjectivity is closely related to biased language, reflecting the point of view of the speaker. Following this intuition, we will firstly seek ways to capture biased language.

We form our basis of claims setting out from argumentation theory, more specifically from the work of Douglas Walton [68]. Walton introduces the dependency between bias and argumentation, mentioning how hard it is to detect bias since it is hidden in the language. He firstly emphasises the two aspects of the term argument saying that "*[f]irst, an argument is a claim with reasons offered to support it but second, an ar-*

gument has more than one side". Then, he points out that what is claimed in an argument is indeed a viewpoint which has two attributes, the proposition (the statement; true or false) and the attitude (the orientation; pro, contra or neutral). Attitudes are conveyed in the statements through some syntactic features and emotive clues or loaded terms which are the words that can cause positive or negative sentiment in the hearer. Walton highlights that excessive or incorrect use of emotive language can cause fallacies in argumentation. Indeed, he states that [68, p. 224]

> The problem with the use of emotive terminology in argumentation is not that the use of such language is inherently wrong. [..] The problem is that putting forward a statement that contains such emotive terminology may conceal the fact that a conclusion is drawn or advocated from the statement and that the statement with the conclusion is an argument.

Hence, it is obvious that emotive terms can propose new statements, bring in new premises in an argument that tries to conclude an issue. Meanwhile, in support of the above quoted proposition, Walton conveys a remark from the book of Johnson and Blair [69], saying that "*[a] loaded term is a label attached to something* ***in a way*** *that makes the statement containing the labeling either debatable or false.*". The expression "in a way" that we emphasized is the key to our investigation. The choice of "that way" in speech can be asserted to be the factor determining if the content is biased or not.

Supporting and formalising his overall conjecture about biased language in argumentation, Walton specifies two key concepts, "innuendo" and "persuasive definitions" as indicators of bias. Innuendo is a method in argumentation that conveys its message through the speech not uttered on the spot but should have been said explicitly. Thus, innuendo might push the audience to make presumptive inferences. Actually, we thought innuendo and implicature detection would provide a good measure in our subjectivity analysis work. However, it would be difficult to extract the necessary features from the text. Although not directly relating to bias and argumentation, there are studies that deal with presumptive expressions; some of which are mentioned by Potts in [70].

The second indicator of bias is given to be the concept of "persuasive definitions" firstly put forward by Charles Stevenson in 1938 [71]. After giving the distinctions between lexical and stipulative definitions of terms in language, Walton describes persuasive definition as [68, p. 247]

> A persuasive definition takes a term that has a conventional lexical meaning in normal usage, and then presents a partly stipulative definition of a kind that support one side and goes against the other side of an issue in a persuasion dialogue.

More compactly, he states that it is the persuasive definitions that attribute a 'spin' on words through making a redefinition in a positive or negative way. Following that, he points out that excessive or incorrect occurrence of persuasive definitions, as well as loaded terms, can cause -even unintended- fallacies in argumentation and, consequently, in a text or a speech, these fallacies can lead to a biased tone.

In a narrower view of persuasive definitions, arguments can be adapted to news stories; persuasion can be considered as the act of manipulation. As both are capable of diverting the topic of discussion through presenting new premises, we deduce that what subjectivity is to news-telling is as what bias is to argumentation. Hence, based on this manipulation-persuasion duality, we conjecture that what makes a piece of news text subjective is the manipulating voice explicitly or implicitly heard over it. In order to capture that manipulating voice, we integrate a redefinition / description measuring mechanism to our subjectivity detection system, inspired by the concept of persuasive definitions.

We consider abstractness of words as an auxiliary dimension of biased tone. A term is called abstract if it cannot be perceived by sensorial organs. In this regard, we claim that significant occurrence of these words in the news texts can lead the reader to a biased and even bounded -notwithstanding, subjective- envisioning of the story through impelling those abstract concepts from the voice of the authority, i.e. the newspaper. In order to lay down a more integrated account of subjectivity, we also take the 'inherently subjective actions (verbs)' as another feature, which are studied

and collected in [72]. However, from this list, we filtered the words in the manner we object to their private state approach such as eliminating the words like think, accept, presume.

All in all, we claim that subjective manner in texts is made up of the significant occurrence of redefinitions and descriptions, together with the significant use of inherently subjective and abstract words. To learn these two features, which are based on the above claims, we apply supervised learning algorithms. Our approach can be applied to texts from domains other than the news medium, as well. The formal account of our feature extraction mechanisms is given in Section 4.1.

2.3.2. How can we automatically detect the subjectivity in text?

In the light of the theoretical basis we formed in the previous subsection, we summarize our claims below, as a foundation of our computational framework. Our approach to detect subjectivity in texts comprises basically the following claims and assumptions:

(i) The whole piece of text (the document) is the discourse unit that carries the property of being subjective or objective. Thus, we examine subjectivity at the document level.

(ii) The subjective manner in texts, in particular in the news texts, is made up of:

- Redefinitions and descriptions in the form of significant use of adjectives that charge nouns (entities) and adverbs that charge nouns and verbs (entities and actions). We identify this property by calculating the number of descriptors (adjectives and adverbs) normalized by the total number of entities and actions, separately.
- Significant use of abstract concepts and subjective actions. We identify this property by calculating the number of abstract nouns and subjective verbs normalized by the total number of words in the text, separately.

We argue that these features are discriminative for the subjectivity property of a whole piece of text. In Chapter 5, we illustrate the validity of this claim empirically.

3. DATASET

3.1. Raw Collection

Our dataset contains news items from three Turkish newspapers, Radikal[1] , Sol-Haber[2] and Vakit[3] , each representing markedly different political viewpoints. The texts are retrieved through crawling the web sites of the papers category by category. As a total there are 66886 texts in our raw collection from seven different categories. Table 3.1 gives the statistics of our collection.

Table 3.1. Collection statistics by source.

	source		
	radikal	*solhaber*	*vakit*
Number of texts	27955	10689	28242
Number of categories	7	10	11
Average number of words per text	283.35	310.79	253.16
Average word length	7.66	7.92	7.85

In Table 3.2, we present the number of texts and average number of words of each source over four categories in our raw collection. Each of these four category implies the same news section for each three source, though their names differ. In Table 3.3, the non-overlapping categories in our collection can be found, together with the number of texts and average number of words per source inside. We determined the temporal range of our corpus by taking those contiguous months that contain the most texts from each source. In the end, the best overlap appeared to be in the range July 2012 - April 2013. Thus, we have built a raw corpus of news texts in this temporal range from three sources, Radikal, SolHaber and Vakit; from the categories world, turkey, economics and politics; containing 59244 texts as a total after excluding the news texts

[1] `http://www.radikal.com.tr/`
[2] `http://haber.sol.org.tr/`
[3] `http://www.habervaktim.com/`

from the non-overlapping categories and dates. Table 3.4 presents the statistics of that selected news collection per category for each source. The texts to be annotated are drawn from this collection to have a distribution of categories and sources over months as balanced as possible. As will be detailed in the next section, we have selected and annotated 1400 texts from this corpus.

Table 3.2. Collection statistics by resource and overlapping categories.

	source								
	radikal			*solhaber*			*vakit*		
	Category name [*eng.*]	Number of texts	Average number of words per text	Category name	Number of texts	Average number of words per text	Category name	Number of texts	Average number of words per text
Overlapping category names	politika [*politics*]	2617	376	devlet ve siyaset [*politics*]	5036	313	siyaset [*politics*]	3821	338
	ekonomi [*economy*]	2891	262	ekonomi [*economy*]	527	379	ekonomi [*economy*]	2104	249
				sonuncu-kavga [*economy*]	915	333			
	dünya [*world*]	4333	181	dünyadan [*world*]	2407	276	dünya [*world*]	6154	167
				enternas-yonal-gundem [*world*]	96	530			
	türkiye [*turkey*]	12520	272	kent-gündemleri [*turkey*]	1216	305	güncel [*turkey*]	14470	262
				soldakiler [*turkey*]	492	314			

Table 3.3. Collection statistics by resource and non-overlapping categories.

	source								
	radikal			*solhaber*			*vakit*		
	Category name [*eng.*]	Number of texts	Average number of words per text	Category name	Number of texts	Average number of words per text	Category name	Number of texts	Average number of words per text
Non-overlapping category names	sinema [*cinema*]	1496	626				kültür-sanat [*culture& arts*]	61	407
	spor [*sports*]	2402	238				spor [*sports*]	110	261
	hayat [*life*]	1738	271				aile-yaşam [*family & life*]	72	331
							bilim [*science*]	266	191
							sağlık [*health*]	944	257
							eğitim [*education*]	300	266
							medya [*the media*]	298	247

Table 3.4. Raw collection's statistics per category by sources.

source	category	number of texts	average number of words per text	average word length
radikal	economy	2889	262.57	7.54
	politics	2612	376.54	7.77
	turkey	12492	273.16	7.72
	world	4333	181.37	7.88
solhaber	economy	1442	349.99	7.8
	politics	5036	313.12	7.92
	turkey	1708	307.41	7.9
	world	2503	285.85	8.03
vakit	economy	2099	249.46	7.67
	politics	3806	339.69	7.86
	turkey	14223	266.06	7.81
	world	6100	168.04	7.99

3.2. The Annotated Corpus

3.2.1. The Need for an Annotated Corpus

There should be a gold-standard dataset to be used, as required by all the machine learning experiments, for training (if supervised) and testing the learning algorithms. While some studies use the categorical attributes of texts as their labels [5, 6], many studies create or use manually-annotated datasets. However, there is no such manually annotated or subjectivity-categorized dataset which is published and used previously in the scope of Turkish language resources [4] . Thus, we created a subjectivity-labelled set of news texts as a part of this study. A subset of our raw collection, which we have

[4]Sentiment annotated sets of Turkish texts are made available in [19,26,59], and in [50], a relatively small set of emotion-annotated fairy tales in Turkish is given.

documented in the previous section, is annotated at the document-level as subjective or objective by human annotators. Before introducing the details of our annotation study, we would like to present some of the annotation schemes developed in the scope of subjectivity / sentiment analysis.

We start the survey with annotation guidelines of the widely known and publicly available subjectivity corpus, Multi-Perspective Question Answering (MPQA), creation of which is detailed in the 2005 study of Wiebe *et al.* [73]. This collection contains 10657 sentences annotated for their attitudes (as subjective or objective) at an intensity scale (medium, high or low) and with the varying details of source and targets of the attitudes in the sentences. For each attribute of subjectivity, three annotators mark the related expressions in the sentences. The annotators are exposed to a 40-hour training for the guidelines of the schema of this study (on what to annotate and how with examples), as well as the technical details of the annotation tool. The lack of formal criteria to determine the subjectivity of texts is emphasized. The task of annotation is reported to have lasted six months while each annotator was working for it 8-12 hours a week.

The authors report the agreement measures of the labelled parts using different metrics since the levels of annotation varies. For example, to study the agreement of annotators on the expressions that they marked for implying private states where the length of marked parts might differ, they define a metric called *agr* to express the rate of agreement between the annotators a and b, similar to f-score, such that $agr(a||b) = \frac{|A \text{ matching } B|}{|A|}$, where A is the set annotated by a and B is the set annotated by b. They indicate that this metric is suitable for the cases when the annotators may mark non-overlapping sets of items. In the end, the authors take the mean of the *agr* values of each three annotator pair and report the result as the inter-annotator agreement of that task. As for the sentence-level subjectivity judgement, since the annotators label the same set of sentences, the authors use Cohen's kappa [74]. The kappa agreement rate is calculated by the formula $\kappa = \frac{P(A)-P(E)}{1-P(E)}$ where $P(A)$ is the observed agreement rate between two annotators and $P(E)$ is the expected (by-chance) agreement rate between the two annotators. The authors take the average of the label-

wise agreements to get the resultant agreement value. Most of the agreement values appear to be above 80%, which they state to be acceptable. The average pairwise kappa value for the sentence-level subjectivity judgement they report is 77%. They include a different treatment of this measure; they say that agreement value rises to 87% after the removal of the sentences which they know to be 'borderline subjective'.

In [75], in order to construct a subjectivity-annotated corpus of sentences in Korean language, Shin *et al.* follow and enrich the approach of Wiebe *et al.* to construct MPQA [73]. They collect 8050 sentences from news articles to be annotated by 3 annotators. In this study, they report the results of the annotation of two 100-sentence sets to show both inter-annotator and intra-annotator agreement results. They apply two annotation rounds and report the results of two agreement measures, Krippendorff's alpha [76] and F-measure (calculated from the averages of pairwise recalls of annotators). Krippendorff's alpha agreement rate is calculated using the formula $\alpha = 1 - \frac{D_o}{D_e}$ where D_o denotes the average of the differences between the values within the annotated units regardless of who assigned them and D_e denotes the average of the differences between all the annotation values regardless of who assigned them.

In their study of polarity classification and related datasets, Veselovská *et al.* create a sentiment-annotated corpus of text segments in Czech language [30]. In order to examine the comparative performances, they use three different resources, namely, movie and product review texts and news texts, respectively. For having the news corpus labelled, they work with two annotators who labelled 410 textual segments for polarity. The authors report that 63% inter-rater agreement value is obtained using the metric Cohen's kappa. They indicate that their annotation instructions might be prone to subjective evaluation of the annotator; so this low agreement can be attributed to the vagueness of the instructions, as well as the relatively higher subjective tone of the news articles in their corpus.

3.2.2. The Annotation Scheme

The general approach in the field is to have a set of textual items annotated by two or three people, calculating the inter-annotator agreement rate using such as kappa or alpha measures [77] in order to give an account about the reliability and validity of the annotated data [76].

Having inadequate resources, we have employed a different and efficient method, suggested by Yu and Hatzivassiloglou [15]. In this scheme, the whole set of items is split into N windows, each having i items. Each window will be assigned to one annotator, such that the process needs N users and each user will annotate i items in total. Meanwhile, each window is divided into two parts, one with size x and the other one with size $i - x$, such that the first part of size x will be shared by the next window. Since each pair (two consecutive windows) shares x items, they will have been annotated by two different people at the end. All in all, the process results in two disjoint sets of annotated textual units, one containing $(N/2) * x$ items labelled by two users and the other one having $N * (i - x)$ items annotated by a single person.

In our case, our unit of text is news items, together with their titles and bodies, and the task is to annotate them by their intensity level of subjectivity. According to the method given in [15], which we described above, we determined N, the number of annotators, to be 70; i, the number of texts a person is to annotate, to be 30 and x, the number of common texts of each two persons, to be 20. Thus, the system outputs 700 twice-annotated texts and 700 single-annotated texts.

We compiled a set of 1400 texts which are selected randomly from our raw collection to form a fair distribution of texts throughout resources, categories and months. Annotation is based on five labels, namely, strongly objective, weakly objective, strongly subjective, weakly subjective and lastly, uncertain. The process is performed online through a dynamic web site created by us. The users are firstly introduced to a guideline that 1) describes our work; 2) explains our definition of subjectivity; 3) presents the subjectivity labels with sample texts and 4) gives suggestions

about some possible caveats regarding 'subjective annotation'.

3.2.3. Labelled Datasets and the Protocol

Our annotation process resulted in three distinct sets[5] of labelled texts, which we named as `corpus1`, `corpus2` and `corpus3`. Below we present their details.

Table 3.5. Label distributions in the three annotated corpora.

dataset	**labels**				**total**
`corpus1`	SS	WS	SO	WO	
	57	139	72	77	345
`corpus2`	$Subj$		Obj		
	288		216		504
`corpus3`	SS	WS	SO	WO	
	135	253	137	174	699

`Corpus1:` This corpus contains 345 texts which received exactly the same label from both annotators. Thus, each text can have one of the subjectivity labels (SS, WS, SO, WO). The number of texts per label in this corpus, together with the other corpora, is listed in Table 3.5.

Table 3.6 presents agreement values for each label. As explained previously, there are 700 twice-annotated texts. There were 35 pairs of annotators and each pair annotated 20 texts together. Therefore, the texts in this set represent two different judgements. However, since not all of the annotators annotated exactly the same set of items, we cannot calculate the state-of-the-art kappa or alpha measures for showing(estimating) the inter-annotator agreement. Instead, we take the ratio of matching evaluations for each label as was done in [15]. More formally, in this dataset D, each text t carries two labels, $label1$ given by the first annotator and $label2$ given by second annotator. Then, A being an arbitrary label:

[5]This collection is freely available and can be asked from the author.

$$agreement(A) = \frac{|\{label1(t) = A \ AND \ label2(t) = A \mid \forall t \in D\}|}{|\{label1(t) = A \ OR \ label2(t) = A \mid \forall t \in D\}|} \tag{3.1}$$

Table 3.6. Agreement values of each label in `corpus1`.

label	**agreement (our scheme)**	**agreement (random annotation)**
strongly subjective (SS)	0.32	*0.13*
weakly subjective (WS)	0.35	*0.09*
strongly objective (SO)	0.33	*0.17*
weakly objective (WO)	0.30	*0.18*
Average	0.32	0.14

The average agreement value in our annotation scheme is 0.32 and the average agreement value for random annotation is 0.14. Though the agreement rate is not very good, this set forms an acceptable annotated dataset. Moreover, agreement values over labels seem stable. We additionally calculated Cohen's kappa for each pair of annotators who annotated the same set of texts; then, we took the mean of those kappa values. The resultant mean is 0.31 (with std 0.29), which is evaluated to be a fair agreement [77]. This low agreement value, as well as fluctuating trend of agreement values that we saw throughout pairs, can be attributed to the quality of the raw data and the approach of the annotators. Firstly, although we tried to select the texts as balanced as possible, the difficulty level of the sets of texts assigned to pairs inevitably differs. Secondly, by intuition, we can say that annotator bias might increase, as the number of annotators get higher given that annotator training was not intensive. We worked with 70 annotators, which is relatively high, directly influencing the agreement rate. The annotation process took one month and it was not feasible to complete it with much smaller numbers of annotators.

`Corpus2:` This corpus is obtained from the set of texts labelled by two annotators such that we took the texts that received the same label class (i.e. subjective or

objective) from both annotators but their intensity (i.e. strong or weak) may not match. Therefore, each text in this set has one of the labels $(Subj, Obj)$. Table 3.5 gives the label distributions. We investigate the effect of intensity mismatches on learning performance and compare these two datasets in this regard (Section 5.2.4).

We calculated the agreement values of the labels in this corpus, as previously we have done for `corpus1`, using the Equation 3.1. The values can be found in Table 3.7. The agreement values are higher than random and being calculated over two classes only, well above those of `corpus1`.

Table 3.7. Agreement values of each label in `corpus2`.

label	**agreement (our scheme)**	**agreement (random annotation)**
subjective ($Subj$)	0.60	0.30
objective (Obj)	0.53	0.34
Average	0.56	0.32

`Corpus3:` The initial set contains 700 texts annotated by only one annotator. Excluding one text which was labelled 'uncertain', this dataset has 699 texts in total. These texts reflect a single annotator's judgement and no agreement study can be done consequently. Table 3.5 shows the label distributions. We collected it to see whether there are any performance differences when we run learning algorithms on this dataset and the previous corpora, namely `corpus1` and `corpus2`, which are filtered by the matching evaluations of two annotators.

4. METHODOLOGY

Given a set of texts, we extract their values on three supersets of features, namely, part of speech (POS) related features, lexicon-based features and the usage of exclamation mark in the title. After mapping the texts onto the feature space with their target class values, we use both binary and multi-class classifiers in a supervised fashion for the task of predicting the subjectivity class of an unknown text.

4.1. Feature Extraction

4.1.1. Part of Speech (POS) Tag Related Features

We prefer to call this class of features 'POS tag related features'. In several studies on subjectivity detection (as we mentioned in Section 2.2), POS tags are assumed to be an indicator of subjectivity and facilitated in classification considerably [14, 15, 23, 50].

As we have previously explained as our approach, we claim that significant use of adjectives that charge nouns (entities) and significant use of adverbs that charge nouns and verbs (entities and actions) indicates subjectivity. We identify this property in two separate measurements by calculating the number of descriptors each (adjectives, adverbs) normalized by the number of entities (nouns) and actions (verbs), as opposed to evaluating the words from these POS tags individually, for example as given in [23]. These word classes are said to be inherently exhibiting sentiment [78, 79].

In what follows, D represents the set of all texts; t represents an arbitrary text as a set of its words w; $S(t)$ represents the set of unique lexical forms of the word w occurring in text t. More formally, $S(t) = \{(root(w), postag(w)) \mid \forall w \in t\}$; that is, $S(t)$ contains root and POS tag of each word occurring in t.

For lemmatization and POS tagging, we used the morphological parser developed by [80]. Sak's morphological analyser [80] outputs a set of morphological parses of a

given word in the order of likelihood. We assumed the first parse to be the most likely one and used it. A parse is made up of the root of the given word and the POS tag of the root, followed by the morphemes, if any, along with their attributes like type and name of the morphemes. In the parses, POS tag of the input word is not explicitly stated; however, it can be inferred using the morphological characteristics of Turkish. Thus, using the properties of the output morphemes, we could extract the POS tag of the input word.

4.1.1.1. Description ratio for entities. This measure, which we call 'entity description ratio *EntityDesc*()', denotes weight of the described entities in a text. It outputs a single value for a text t by taking the number of its adjectives (descriptors) normalized by the number of nouns (entities) in the text. More formally, it finds for any text t in D:

$$EntityDesc(t) = \frac{|\{postag = ADJECTIVE \mid \ \forall(root, postag) \in S(t)\}|}{|\{postag = NOUN \mid \forall(root, postag) \in S(t)\}|} \tag{4.1}$$

4.1.1.2. Description ratio for actions. This measure, which we call 'action description ratio *ActionDesc*()', outputs a single value for a text t by finding the number of its adverbs (descriptors) normalized by the total count of its verbs (actions) and adjectives. More formally, it finds for any text t in D:

$$ActionDesc(t) = \frac{|\{postag = ADVERB \mid \ \forall(root, postag) \in S(t)\}|}{|\{postag = VERB \ or \ postag = NOUN \mid \ \forall(root, postag) \in S(t)\}|} \tag{4.2}$$

Both measures, namely entity and action description ratios, are calculated for both headline and body of the news texts. A comparison of the performance of these

measures can be found in Section 5.2.2.1.

4.1.2. Lexicon-based Features

We made use of two lexicons for subjectivity, where we included only nouns and verbs. Both lists are available in Appendix A.

4.1.2.1. Abstract nouns. We assume that not all but a good deal of abstract words are inherently subjective, so they contribute to the subjective tone of a text. This set has only nouns which denote abstract concepts, 233 words in total. We manually compiled this lexicon selectively out of the list given in [81], which contains 600 Turkish words evaluated for their imagination, abstractness and frequency values.

4.1.2.2. Subjective verbs. This set has only verbs which we assume to denote subjective actions. We think that the occurrence of subjective actions renders subjective voice in the content. The list of verbs is selectively taken from the 'Subjectivity Lexicon' [72] and translated into Turkish, resulting in a total of 266 verbs.

In order to evaluate each of these keyword lexicons introduced above, we take the rate of the intersecting words. Let D represent the set of all texts in the corpus; $W(t)$ represent the set of all the stemmed words of text t and L represent the set of keywords, i.e., the lexicon. To represent this feature set, we calculate the sum of the number of common words (by summing up their total number of occurrences) of lexicon and the text, normalized by the total number of words in the text; that is, for a single text, this measure produces a single value, $KeywordRatio(t)$, such that for any text t in D,

$$KeywordRatio(t) = \frac{|L \cap W(t)|}{|W(t)|} \tag{4.3}$$

We calculated this measure per lexicon, for both the body and the headline

of the news texts, separately. Comparison of this measure with the state-of-the-art alternatives can be found in Section 5.2.2.2.

4.1.3. Usage of Exclamation Mark

We added a binary feature to our system, assuming that an exclamation mark in the news headline contributes to the subjective manner of the text. We could have applied this to the body of the text as well but it could introduce noise and we observed that exclamation mark is used generally in the titles of the news. For a news text n, this measure outputs a single value which is 1 if the headline of n contains exclamation mark ('!'), and is 0, otherwise.

4.2. Classification Approach

4.2.1. Classifiers

We have chosen Support Vector Machines [82] and Naive Bayes algorithms [83] as our supervised classifiers, having known that in the subjectivity and sentiment analysis background work, they are reported to performed well. We briefly present the theoretical basis of these two algorithms. In this section, a training dataset is represented by $\mathcal{D} = \{(\mathbf{x_i}, y_i)\}|_{i=1}^{N}$ where $\mathbf{x}_i$ is the ith data point such that $\mathbf{x}_i = [x_{i1}, ..., x_{it}]$ with x_{ij} being the value of the jth feature and y_i is the true class value of the ith data point such that $y_i \in K$ with $K = \{c_1, ..., c_k\}$ being the set of classes.

4.2.1.1. Support Vector Machines. The Support Vector Machine (SVM) is the general name of a supervised binary classifier, which separates the input space using a decision surface determined by the points called 'support vectors', a subset of the input data points [83]. The decision function of the classifier discriminates the two classes by trying to maximize the margin defining the optimal distance from the decision surface to the other data points. Letting the decision surface be a hyperplane represented by

$\langle \mathbf{w}, \mathbf{x} \rangle + b$ and $y_i \in \{-1, +1\}$, the decision function is basically

$$f(\mathbf{x}) = sign(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{4.4}$$

where x is any data point; $\mathbf{w}$ is the weight vector perpendicular to the hyperplane and $f(\mathbf{x}) \in \{-1, +1\}$, denoting the predicted class of x, under the constraint that each x should satisfy $\langle \mathbf{w}, \mathbf{x} \rangle = -b$. Letting $y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b)$ denote the position of any data point x_i with respect to the hyperplane, SVM requires that $y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b) \geq 1$. Another requirement is that all the data points be as quantitatively far away from the hyperplane as possible for the sake of generalization. More formally, SVM tries to find $\mathbf{w}$ and b such that $y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b) \geq 1 \ \ \forall(\mathbf{x_i}, y_i) \in \mathcal{D}$ and that $\frac{||\mathbf{w}||}{2}$ will be minimum. Finally, with the introduction of Lagrange multipliers λ_i, indicating if x_i is a support vector, the decision function is

$$f(\mathbf{x}) = sign(y_i \lambda_i \langle \mathbf{x_i}, \mathbf{x} \rangle + b) \tag{4.5}$$

If the misclassification errors are to be penalized (due mostly to the data space's being not linearly separable) using the slack variables $\xi_i > 0$ and the regularization parameter C, the constraints are a bit modified such that $y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b) \geq 1 - \xi_i \ \ \forall(\mathbf{x_i}, y_i) \in \mathcal{D}$ that $\frac{||\mathbf{w}||}{2} + C \sum_i \xi_i$ will be minimum.

If this classifier with the linearity assumption is incapable of separating the input space, with the help of a kernel function K, the data is mapped onto a higher dimensional space that allows linear separation. Then, SVM decision function becomes,

$$f(\mathbf{x}) = sign(\sum_i \lambda_i y_i K(\mathbf{x_i}, \mathbf{x}) + b) \tag{4.6}$$

In this work, we used the kernels [84],

- Radial Basis Function (RBF): $K(\mathbf{x}, \mathbf{x}') = \mathrm{e}^{-\gamma||\mathbf{x}-\mathbf{x}'||^2}$
- Linear: $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$
- Polynomial: $K(\mathbf{x}, \mathbf{x}') = (\gamma \langle \mathbf{x}, \mathbf{x}' \rangle + r)^d$
- Sigmoid: $K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \langle \mathbf{x}, \mathbf{x}' \rangle + r)$

4.2.1.2. Multinomial Naive Bayes. Naive Bayes classification is based on the naive assumption that the features of a data point are independent. The decision function of a naive Bayes classifier works by maximizing the conditional $P(c|\mathbf{x})$ denoting the probability of the class c given the text instance or the data point x; in other words, it is the probability that $\mathbf{x}$ is in class c. By Bayes' rule,

$$P(c|\mathbf{x}) = \frac{P(c)P(x_1, ..., x_k|c)}{P(x_1, ..., x_k)}$$

Under the independence assumption, $P(x_i|x_1, .., x_{i-1}, x_{i+1}, .., x_k) = P(x_i|c)$. Omitting the denominator being independent of the class, then,

$$P(c|\mathbf{x}) \propto P(c) \prod_{i=1}^{k} P(x_i|c)$$

Finally, the decision function outputs the most likely class such that

$$f(\mathbf{x}) = \underset{c}{\operatorname{argmax}}\ \hat{P}(c|\mathbf{x}) = \underset{c}{\operatorname{argmax}}\ \hat{P}(c) \prod_{i=1}^{k} \hat{P}(x_i|c) \tag{4.7}$$

In Multinomial Naive Bayes, the input data is assumed to be from a multinomial distribution. The estimators of the parameters $P(c)$ and $P(x_i|c)$ are defined to be [84],

$$\hat{P}(c) = \frac{|\{(\mathbf{x}_j, y_j = c) \mid \forall (\mathbf{x}_j, y_j) \in \mathcal{D}\}|}{|\mathcal{D}|}$$

and

$$\hat{P}(x_i|c) = \frac{|\{(\mathbf{x}_j, y_j) \mid x_i \in \mathbf{x}_j \ \forall(\mathbf{x}_i, y_i) \in \mathcal{D}\}| + \tau}{|\{x_{j,i} \mid x_{j,i} \in \mathbf{x}_j \text{ and } x_{j,i} \neq x_{j,i+1} \ \forall(\mathbf{x}_j = [x_{j,1}, ..., x_{j,k}], y_j) \in \mathcal{D}\}| + \tau k}$$

where τ is the smoothing parameter and k is the dimension of the any data vector $\mathbf{x}_i$, i.e., the number of features in the model.

4.2.2. Subjectivity Classes

As given in Section 3, each text in our corpus can have one of the labels (SS, WS, SO, WO) where SS denotes strongly subjective, WS denotes weakly subjective, SO denotes strongly objective and WO denotes weakly objective. Distribution of these classes per corpora has been given in Table 3.5. Our aim is to examine the effect of the intensity of these labels on learning performance. For example, we expect that the algorithm will give good results when run on the set of strongly subjective and strongly objective texts since these texts are expected stronger and more apparent expressions in terms of the existence and absence of subjectivity, respectively. We also would like to investigate as to whether there will be any performance differences between using the set of subjective and objective texts (disregarding the intensity) and using the other binary-class sets that have texts by their intensity.

In this regard, we have combined the labelled texts in six different ways, each leading to a new dataset, the details of which are given below. Our classifiers are experimented on each of these datasets.

(i) *Original labels / four-class:* In this approach, we take all the texts with their original labels. Each text has four labels and on this dataset, multi-class classification will be performed. We expect to get the smallest accuracy on this dataset since multi-class classification is harder than binary classification and also distinguishing intensity of the subjectivity labels will not be easy as we saw in the annotation phase.

(ii) *Subjective vs Objective / binary-class:* This approach combines all the weakly or strongly subjective texts in the subjective class ($SS \cup WS$) and combines all the weakly or strongly objective texts in the objective class ($SO \cup WO$).

(iii) *Strongly Subjective vs Objective / binary-class:* This combination takes only the strongly subjective texts in the class subjective (SS) and gathers all the weakly or strongly objective texts in the objective class ($SO \cup WO$).

(iv) *Subjective vs Strongly Objective / binary-class:* This combination has all the strongly or weakly subjective texts in the class subjective ($SS \cup WS$) and has only the strongly objective texts in the objective class (SO).

(v) *Strongly Subjective vs Strongly Objective / binary-class:* This combination has only the strongly subjective texts in the class subjective (SS) and has only the strongly objective texts in the objective class (SO).

(vi) *Weakly Subjective vs Weakly Objective / binary-class:* This combination has only the weakly subjective texts in the class subjective (WS) and has only the weakly objective texts in the objective class (WO).

5. EXPERIMENTS

5.1. Setting

The texts are preprocessed to eliminate stopwords and punctuation (except exclamation mark in the title). As explained in Section 4.1, our feature extraction mechanism outputs data matrices with nine feature dimensions such that the POS tag related features class contributes with four dimensions; the lexicon-based features class contributes again with four dimensions and lastly the usage of exclamation mark has one dimension.

In order to learn the mapping of these features, we applied supervised learning algorithms, namely, Support Vector Machines (SVM) and naive Bayes; investigating their effect with different parameters in order to choose the best model. We used multinomial naive Bayes classifier in the experiments, which we ran for the purpose of comparison (see Section 5.2.2.2 for the details), where the feature extractors model the texts as word vectors.

We used one of the machine learning libraries of Python, called scikit-learn [84]. We applied grid search for selecting SVM's parameters. We held SVM experiments with the kernels sigmoid, radial basis function (RBF), linear and polynomial. For RBF, linear and polynomial kernels, we used the penalty values $\{1, 10, 100, 1000\}$. On the polynomial kernel, the degree values $\{2, 3, 4, 5\}$ were tested. According to scikit, the default penalty value is the number of data points to learn and the default degree value is taken to be 3. For RBF, degree is by default equal to 2 in the tool and indeed it has been suggested in [85] to take degree to be 2. With all the parameter combinations, from among 25 models. As for multinomial naive Bayes, we took its smoothing parameter to be 1.0.

All our experiments, namely the central experiment, feature study and ablation study, are run on `corpus1`, which contains twice-annotated texts with coarse-grained

subjectivity labels. We have repeated our central experiment on the other datasets (`corpus2` and `corpus3`) as well, to make a cross-corpus comparison. All results are obtained with 5-fold cross-validation.

5.2. Results & Discussion

We investigated the following questions:

- Does the description ratio perform comparably with the state-of-the-art POS tag related feature measures?
- Do different measures of features affect learning performance?
- Which group of labels results in the best performance? Is the strength (intensity) of subjectivity well-distinguishable?
- Does the agreement level of annotation affect performance? Which corpus, namely `corpus1`, `corpus2` or `corpus3`, gives the best performance?

Performance of each classification experiment is evaluated with accuracy measure, which is the number of correctly classified items normalized by the total number of items in the dataset.

5.2.1. Central Experiment

For extracting features in this experiment, we have applied the description ratio measure for the POS tag related features class and the rate measure for the lexicon-based features class. Our aim is to report the performance of our central model, to be later compared with other feature models as well. Additionally, the setting is applied on each of the six datasets.

Table 5.1 shows the results of this experiment using Support Vector Machine with RBF kernel (SVM). We employed two settings for baseline; the first one is a random classifier and the second one uses the bag of words (BoW) representation of texts as features and uses SVM as classifier. All the six datasets (label unions)

Table 5.1. Cross-validation results of the central experiment and the two baseline settings.

			Accuracy (% (±*std*))		
	label union name	**dataset size**	**Random classification**	**BoW features(SVM)**	**Central features (SVM)**
1	*Original labels (4-class)*	345	23.77 (±5.75)	40.75 (±7.26)	45.28 (±2.98)
2	*Subj* vs *Obj*	345	46.04 (±11.98)	53.21 (±4.88)	68.30 (±5.88)
3	*SS* vs *Obj*	206	48.75 (±10.73)	73.75 (±1.71)	68.75 (±3.83)
4	*Subj* vs *SO*	268	52.68 (±3.70)	73.66 (±1.09)	**79.51 (±5.62)**
5	*SS* vs *SO*	129	52.00 (±13.51)	58.00 (±2.74)	71.00 (±2.24)
6	*WS* vs *WO*	216	53.94 (±14.12)	63.64 (±0.00)	64.85 (±4.60)

have two target classes, except the one shown at row #1 in the table, which has four classes. According to the table, the accuracy of our model on all the six datasets (label unions/combinations) are well above random classification. We see that our model is outperformed by the second baseline featuring BoW representation in only one case that uses the dataset '*SS* vs *Obj*' (row #3).

The dataset '*Subj* vs *SO*' in row #4 has given the highest accuracy. The least performing case is the combination that has the texts with their original labels, shown in row #1. Inherent difficulty of four-label classification is the reasons why this set performs the worst [6] . Among other cases, we see that the dataset '*WS* vs *WO*' performs remarkably low. This dataset has only the weakly-intense texts. We infer that these texts carry similar values; thus, it is not easy for our classifier to distinguish weakly subjective texts from the weakly objective texts. Moreover, the accuracy notably increases when the objective class contains only the texts carrying the 'strong' intensity, comparing the rows #2 to #4 and the rows #3 to #5. We conjecture that the 'weakly objective' texts in our annotated collection are at the border of being subjective and

[6]Lehrman *et al.* report a similar result about binary classification versus multi-level(intensity-level) classification [14].

objective; thus, they might have been treated to be noise by the classifier. We think that the lack of sharper distinctions between subjective and objective expressions is the key to that bottleneck. Removing 'weakly subjective' texts from the set '*Subj* vs *SO*' causes an obvious decrease in accuracy. At the same time, the dataset size is cut half and the majority of the texts are subjective. This might be one of the reasons for this decrease. Another reason should be that 'weakly subjective' texts are already well separable from 'strongly objective' texts. In fact, though not shown in the table, we have performed an extra experiment to distinguish the texts in these two classes. The accuracy was obtained as 73.75%, which is relatively good and supports our hypothesis. However, when we remove the 'weakly subjective' texts from the dataset '*Subj* vs *Obj*', the accuracy is not significantly affected omitting the small increase (see rows #2 and #3).

Overall, our largest accuracy value is 79.51% on the dataset '*Subj* vs *SO*', which gathers all the subjective and only the strongly objective texts. We can say that our feature extraction mechanism gives a good performance once the ground-truth labels are reliable. In the next section, we study the alternative feature measures in order to investigate the possible performance differences.

5.2.2. Comparing Different Feature Extraction Mechanisms

5.2.2.1. Comparing POS tag related feature measures. We have run another experiment using the count measure for POS tag related feature group as an alternative to our description measure, which was explained in Section 4.1.1. Thus, in this experiment, we have taken the count of adjectives instead of rating the number of adjectives to nouns (Equation 4.1); and the count of adverbs instead of rating the number of adverbs to the sum of verb and noun counts (Equation 4.2). The two measures are calculated for both title and body of the news text separately as before.

Table 5.2 shows the accuracy values of this experiment. There are only two cases, namely the datasets '*SS* vs *Obj*' and '*SS* vs *SO*', where the count measure, denoted by COUNT in the table, outperforms our description ratio. Seeing that even only counting

adjectives and adverbs causes such an increase in accuracy, our first inference is that they are an important factor for distinguishing strongly subjective texts. Nevertheless, our description ratio measure appears to give no worse results than the count measure, as well as having smaller variance. The description ratio measure is more context-sensitive than counting POS tags and is intuitively more consistent and correct for detecting subjectivity. Yet it needs a deeper consideration and improvement using for instance phrase-level information. Looking at the overall results, our hypothesis that description ratio measure is a good candidate for being a subjectivity cue has been justified as it is comparable to its alternative.

Table 5.2. Learning accuracy with two feature extraction schemes for POS tag related features, namely, COUNT that uses count measure and our central experiment that uses description measure.

		Accuracy (% (±*std*))	
	label union name	**COUNT**	**Central experiment**
1	*Original labels (4-class)*	42.46 (±2.15)	45.28 (±2.98)
2	*Subj* vs *Obj*	66.42 (±6.17)	68.30 (±5.88)
3	*SS* vs *Obj*	73.13 (±5.23)	68.75 (±3.83)
4	*Subj* vs *SO*	**76.59** (±7.03)	**79.51** (±5.62)
5	*SS* vs *SO*	76.00 (±5.48)	71.0 (±2.24)
6	*WS* vs *WO*	64.85 (±5.07)	64.85 (±4.60)

5.2.2.2. Comparing lexicon related feature measures. Leaving the values of the POS tag related features unchanged as the previous central experiment, we have run two more experiments, using binary presence and tf*idf (product of term frequency and inverse document frequency of a text) measures to calculate the lexicon-based features, respectively. Both of these measures are explained below. Letting L represent the lexicon, $W(t)$ represent the set of lemmata of text t and D represent the dataset, more formally,

(i) Binary presences of the keywords:

This measure produces a vector $\mathbf{b}$ of size $1 \times |L|$ such that $\mathbf{b}[i] = 1$ if $w_i \in W(t)$ and $\mathbf{b}[i] = 0$, otherwise for $w_i \in L$.

(ii) Tfidf weights of the keywords:

This measure calculates the state-of-the-art term weighting measure, tf*idf, [83, 86] which is based on the idea that a word is important for a text if it is frequent in the text and rare in the overall dataset. For a single text t, this measure outputs a vector $\mathbf{v}$ of size $1 \times |L|$ where the ith entry represents the weight of the keyword w_i inside text t, for $w_i \in L$; more formally, $\mathbf{v}[i] = (|\{w = w_i \mid \forall w \in W(t)\}|) * (\log(\frac{|D|}{|\{w_i \in d_j | \forall d_j \in D\}|}))$.

While our keyword ratio measure (Equation 4.3) outputs a single value for a single text, both of these measures produce vectors of length equal to the lexicon size; thus, they increase the dimension of the feature space. We have run two classifiers, namely SVM and multinomial naive Bayes, on these settings. As indicated in Section 4.2.1, naive Bayes classification is said to be suitable for vector space models; in this regard, we investigate its performance in comparison with SVM. The tables 5.3 and 5.4 present the cross-validated accuracy results of the experiments with these two measures as well as our central experiment for comparison, using the classifiers SVM and multinomial naive Bayes, respectively. In both tables, PRES maps the keywords for their binary presence in the text to vectors and TFIDF vectorizes keywords for their tf*idf weights in the text and our central experiment takes the total number of the present keywords divided by the total number of terms in text for lexicon-based features.

In Table 5.3, it appears that the central experiment that uses keyword ratio measure for lexicon-based feature group results in generally larger accuracy than tf*idf (TFIDF) and binary presence (PRES) weighting using SVM. Additionally, PRES and TFIDF have similar accuracy results in all cases. We can attribute this to the individual words' having similar vector representations and the individual tf*idf values being close to one. Thus, distinguishing subjectivity implying words from the others can be expected to be more difficult [7] . Interestingly, these two measures has given relatively

[7] We also experimented with the frequency measure and obtained nearly the same results.

Table 5.3. Learning accuracy with three feature extraction schemes using SVM algorithm.

		Accuracy (% (±*std*))		
	label union name	**PRES**	**TFIDF**	**Central experiment**
1	*Original labels (4-class)*	39.62 (±0.00)	39.62 (±0.00)	45.28 (±2.98)
2	*Subj* vs *Obj*	58.11 (±6.98)	61.89 (±10.95)	68.30 (±5.88)
3	*SS* vs *Obj*	**73.75** (±1.71)	**73.75** (±1.71)	68.75 (±3.83)
4	*Subj* vs *SO*	73.17 (±0.00)	73.17 (±0.00)	**79.51** (±5.62)
5	*SS* vs *SO*	59.00 (±4.18)	57.00 (±2.74)	71.0 (±2.24)
6	*WS* vs *WO*	63.64 (±0.00)	63.64 (±0.00)	64.85 (±4.60)

higher results with the datasets '*SS* vs *Obj*' and '*Subj* vs *SO*', where only one of the classes has strongly-intense texts. We think that the vector space approach followed in PRES and TFIDF is more sensitive to noise caused by weakly-intense texts; indeed, these two measures are less capable of distinguishing weakly subjective texts than our lexicon rating measure. Accuracy values of PRES and TFIDF are higher than ours in only one case (row # 3) where the dataset is '*SS* vs *Obj*'. It appears that taking the individual weights of the keywords in objective and strongly subjective texts is more informative than taking their sum total. Still, we can say that lexicon rating has given relatively better results than the individual weighting of lexicon words and it is computationally much cheaper. Moreover, it appears that the keywords in our lexicons are a promising factor for detecting subjectivity; though experimenting with different lexicons as well will show the effect of lexicon choice more precisely.

In Table 5.4, we present the results of the experiments where we applied multinomial naive Bayes algorithm for classification, employing respectively PRES and TFIDF measure for lexicon-based features. It appears that the accuracy values of TFIDF slightly exceed those of PRES, which is in line with the results given in [38]. In this setting, both mechanisms give their best result on the dataset '*SS* vs *Obj*'. Additionally, standard deviation of accuracy values across folds are remarkably high except

the case 'WS vs WO' (row #6) and the four-class classification case (row #1) where the performance is significantly lower than the others. We infer from these results that the vector-space modelling in our scope is sensitive to annotation quality and the differences of texts within classes.

Table 5.4. Learning accuracy with three vector-space model feature extraction schemes using multinomial naive Bayes algorithm.

		Accuracy (% (±*std*))	
	label union name	**PRES**	**TFIDF**
1	*Original labels (4-class)*	40.00 (±1.58)	40.75 (±1.69)
2	*Subj* vs *Obj*	56.23 (±3.63)	58.11 (±4.50)
3	*SS* vs *Obj*	**76.88** (±2.80)	**79.38** (±4.19)
4	*Subj* vs *SO*	73.51 (±6.79)	73.14 (±7.69)
5	*SS* vs *SO*	75.00 (±11.73)	75.00 (±10.00)
6	*WS* vs *WO*	63.03 (±1.36)	63.03 (±1.36)

5.2.3. Ablation Study

5.2.3.1. Examining the Features. We run an ablation study to examine the effect of features by running experiments on datasets where each time one feature group or a dimension is discarded. The results can be seen in Table 5.5.

(i) *The effect of features individually:*

- Discarding the entity description ratio feature (ADJ) causes a relatively larger decrease in accuracy than the action description ratio (ADV), especially in the label combinations '*Subj* vs *Obj*' and '*WS* vs *WO*'.
- The exclusion of subjective verbs (SUBJ) from our lexicons causes a slight increase in accuracy with the datasets '*SS* vs *Obj*' and '*SS* vs *SO*'. We infer that this lexicon does not help much in distinguishing strongly subjective texts. In other cases, we see that accuracy values have slightly decreased.

Table 5.5. Learning accuracy on discarding features individually and in groups.

			Accuracy (% ($\pm std$))						
			discarding individual features					**discarding feature groups**	
	label union name	**full set**	**ADJ**	**ADV**	**SUBJ**	**ABS**	**EXCL**	**Lexicons**	**POStags**
1	*Original labels (4-class)*	45.28 (±2.98)	43.40 (±1.89)	44.91 (±3.63)	43.77 (±3.63)	40.75 (±4.54)	43.77 (±2.80)	38.87 (±3.43)	47.17 (±3.53)
2	*Subj* vs *Obj*	68.30 (±5.88)	64.53 (±5.06)	67.92 (±5.17)	66.42 (±2.07)	64.91 (±2.15)	67.55 (±5.4)	64.91 (±3.68)	66.79 (±5.60)
3	*SS* vs *Obj*	68.75 (±3.83)	68.75 (±3.83)	68.75 (±3.83)	71.88 (±3.83)	69.38 (±4.07)	69.38 (±4.07)	73.12 (±3.56)	68.75 (±3.83)
4	*Subj* vs *SO*	79.51 (±5.62)	77.07 (±6.12)	76.59 (±6.59)	79.02 (±3.27)	77.56 (±4.69)	79.51 (±6.12)	76.10 (±3.18)	71.71 (±7.44)
5	*SS* vs *SO*	71.00 (±2.24)	71.00 (±2.24)	70.00 (±6.12)	73.00 (±4.47)	70.00 (±7.07)	67.00 (±8.37)	71.00 (±9.62)	69.00 (±5.48)
6	*WS* vs *WO*	64.85 (±4.60)	61.21 (±1.36)	63.64 (±3.03)	63.03 (±3.32)	66.06 (±1.36)	66.06 (±3.95)	64.24 (±2.54)	63.64 (±4.29)

However, removal of abstract nouns (ABS) from the lexicons causes a more apparent decrease in accuracy, in all cases except 'WS vs WO'. It can be said that abstract nouns are more effective for subjectivity classification than subjective verbs. Indeed, Dias *et al.* report that the abstractness of words are the most effective feature among the features using POS tags and affective words for supervised subjectivity learning [87].

- When we discard the feature for the use of exclamation mark in the headline of the text (EXCL), we observe no significant change in performance.
- Overall, we can say that the entity description ratio and abstract nouns rating are the most effective features in capturing subjectivity.

(ii) *The effect of features in groups:* We have investigated the effect of the exclusion of feature tuples, namely, the lexicon based features group with two distinct dimensions for subjective verbs and abstract nouns, denoted by 'Lexicons' in the table; and the POS tag related features group with two distinct dimensions for entities and actions, denoted by 'POStags'. It appears that the POS tag related features has more influence in subjectivity detection as their absence results in the sharpest decrease, which can be seen in row #4 where the dataset is '*Subj* vs *SO*'.

5.2.3.2. Examining the Textual Structure. We have run two experiments where one uses only the content (body) of the news text (excluding the headline) and the other one uses only the headline(title) of the text discarding the content. We examine as to whether the title of the text contributes to prediction performance.

In Table 5.6, we present the accuracy results of the two experiments. We would like to note at this point that as we have explained in Section 4.1, we make use of both the headline and the body of the text, calculating all the features for both parts. Thus, exclusion of one part results in removing the half of the feature values, decreasing the size of the feature space. Inspecting the results of the first experiment, we see that the exclusion of the headline results in an apparent decrease in accuracy in all the datasets, except 'SS vs Obj' (row # 3). Indeed, the accuracy have increased on the

removal of the body as well (see next column). We infer that in this dataset, the titles and bodies of the texts do not present consistent values and introduce noise for one another. Compared to the accuracy of the dataset '*SS* vs *SO*', we can say that for the weakly objective texts, the absence of headline increases accuracy.

As for the exclusion of the body, we again see that it causes a remarkable decrease in accuracy for all the datasets, except '*SS* vs *Obj*'. For the case '*SS* vs *SO*' (row # 5), we see that the headline of the text is slightly more important than the body. The highest decrease in accuracy occurs in the dataset '*Subj* vs *Obj*' with very small deviation in accuracy across folds as well. Overall, we infer that the effect of headlines is important for predicting subjectivity.

Table 5.6. Learning accuracy values on including only the body and only the title of the news text, respectively.

			Accuracy (% (±*std*))	
	label union name	**full text**	**using only BODY**	**using only HEADLINE**
1	*Original labels (4-class)*	45.28 (±2.98)	44.91 (±5.57)	38.49 (±1.69)
2	*Subj* vs *Obj*	68.30 (±5.88)	67.92 (±4.22)	56.23 (±0.84)
3	*SS* vs *Obj*	68.75 (±3.83)	71.25 (±2.61)	70.62 (±8.15)
4	*Subj* vs *SO*	79.51 (±5.62)	74.63 (±6.81)	73.17 (±1.78)
5	*SS* vs *SO*	71.00 (±2.24)	66.00 (±7.42)	68.00 (±7.58)
6	*WS* vs *WO*	64.85 (±4.60)	64.24 (±4.49)	63.03 (±1.36)

5.2.4. Comparing Different Annotation Types

As given in Section 3.2.3, we have three different corpora, namely `corpus1`, `corpus2` and lastly `corpus3`, each having a different annotation protocol. We investigate whether or not (or to what extend) the classifier better performs when run on the corpus that reflects the matching judgements of two annotators than being run

on the corpus reflecting single judgements. And secondly, we ask if filtering the texts by only the matching subjectivity classes (subjective / objective) or by the matching subjectivity class and intensity (strong / weak) would result in any difference in performance.

We used only the label combination '*Subj* vs *Obj*' since `corpus2` has only these two classes. We have trained our central model on the `corpus2` and `corpus3` and tested it on the texts from `corpus1`. We run two experiments, in the first one whose results given in Table 5.7, the training sets contain all the texts per corpus, thus, dataset sizes are unequal and in the second one given in Table 5.8, the training set sizes are held equal.

In Table 5.7, we see that the performance of the classifier for each corpus does not change remarkably; though, we see that the best accuracy is attained when training set is from `corpus1` despite its having the smallest size. We would like to note that `corpus2` actually has 504 texts; however, after the removal of the test items from its full-size training set, it remained to be a set with 451 texts only. The performance on `corpus3` is the lowest although it has the largest size. We infer that the possible inconsistency in single-annotator judgements is reflected in the prediction performance as well.

Table 5.7. Accuracy results of our model trained on three corpora using all their texts and tested on `corpus1`.

corpus name	**training set size**	**accuracy (SVM)**
`corpus1`	292	68.30 (±5.88)
`corpus2`	451	67.92 (±4.62)
`corpus3`	699	66.04 (±4.62)

In Table 5.8, we present the results of another experiment where we held all

the training set sizes equal and again took the test items from `corpus1`. All sizes being equal, the classifier trained on `corpus1` gives the best performance. We infer that intensity match of labels, therefore the reliability of the corpus, causes an obvious increase in accuracy. Reflecting the agreed viewpoint of two persons, rather than a single viewpoint and not containing intensity disagreement on labels, the texts in `corpus1` embody more consistent values of features, compared to those of `corpus2` and `corpus3`.

Table 5.8. Accuracy results of our model trained on three corpora using all their texts and tested on `corpus1`.

corpus name	**training set size**	**accuracy (SVM)**
`corpus1`	292	68.30 (±5.88)
`corpus2`	292	65.66 (±7.11)
`corpus3`	292	64.15 (±7.55)

5.2.5. Visualizing the Predictions

To interpret our prediction results, we visualized the texts using their predicted labels, their categories and their sources. Predictions are the outputs of one of the five folds of our best performing test, which is the classification experiment where the dataset is *Subj* vs *SO* with the classifier SVM (See Table 5.1 for the details.). In its test set, there are 41 text instances, 30 of which are predicted to be subjective and 11 of which are predicted to be objective. Below in Table 5.9 we give the distribution of news sources and categories in this set over labels.

We present a baseline visualization, which can be developed as a future work to make it more comprehensible for the purpose of examining the prediction results comparatively with the real-world knowledge in the scope of subjectivity features and media analysis.

Table 5.9. The distribution of news categories and sources over the predicted subjectivity classes in the test set.

	source			**category**			
	radikal	*solhaber*	*vakit*	*politics*	*world*	*turkey*	*economy*
subjective	6	12	12	9	9	5	7
objective	3	1	7	0	5	3	3

In Figure 5.1, we see that all the categories except *Politics* have the same subjective-objective distribution, containing texts that are predicted to be subjective around 65%. All the texts in the category *Politics* are predicted to be subjective.

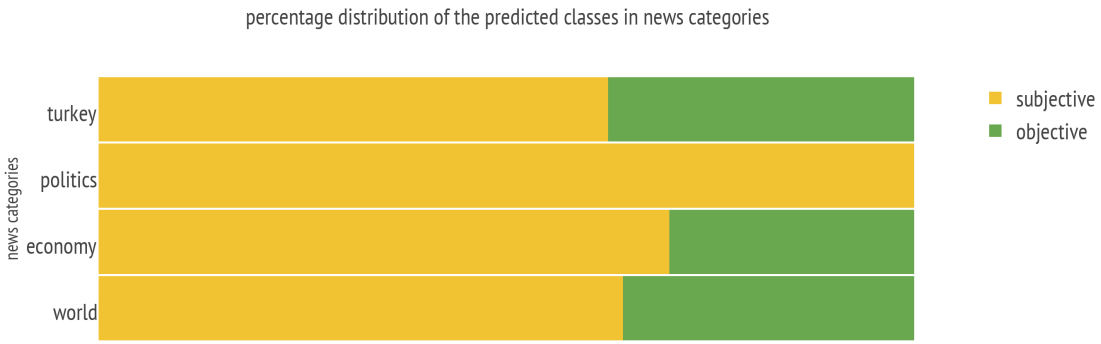

Figure 5.1. Percentage distribution of the predicted classes in news categories.

Figure 5.2 is actually another version of Figure 5.1 above. In this figure, we see that all the categories have nearly equal distributions of texts in both of the subjectivity classes. There is no text from the *Politics* category in the objective class, as we have seen above.

We also give the counterparts of the above figures for the sources of news texts. In Figure 5.3, we see that majority (92%) of the texts of the source *SolHaber* are predicted to be subjective. The other sources have around 65% of their texts predicted to be subjective.

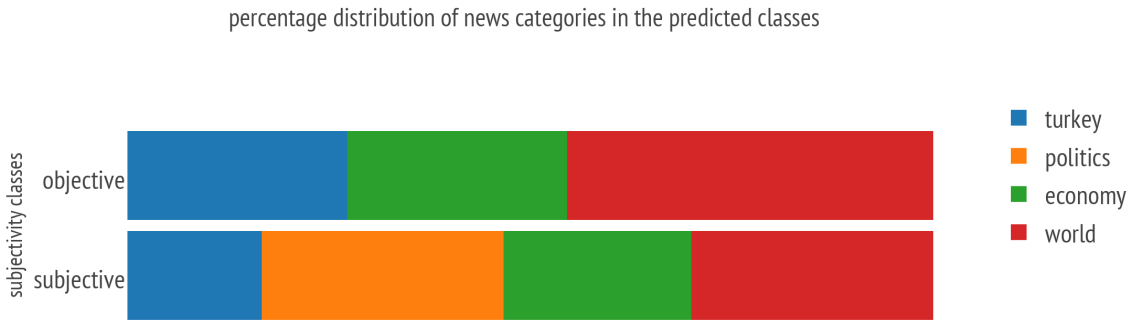

Figure 5.2. Percentage distribution of news categories in the predicted classes.

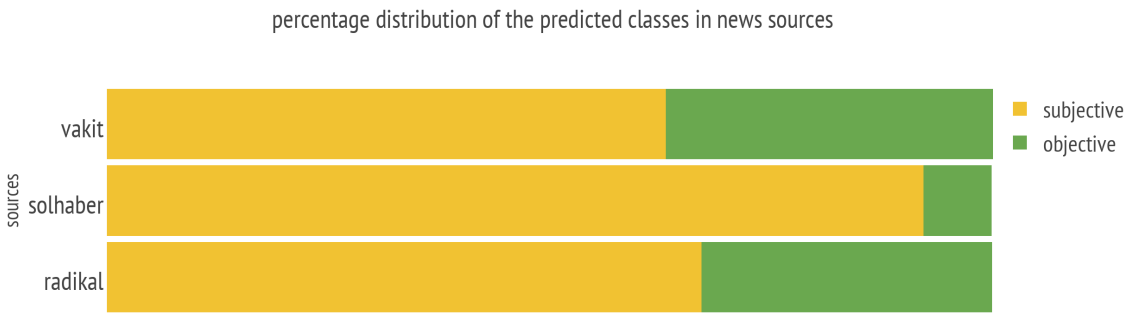

Figure 5.3. Percentage distribution of the predicted classes in news sources.

According to Figure 5.4, the newspaper *Vakit* has more objective texts compared to the sources *Radikal* and *SolHaber*.

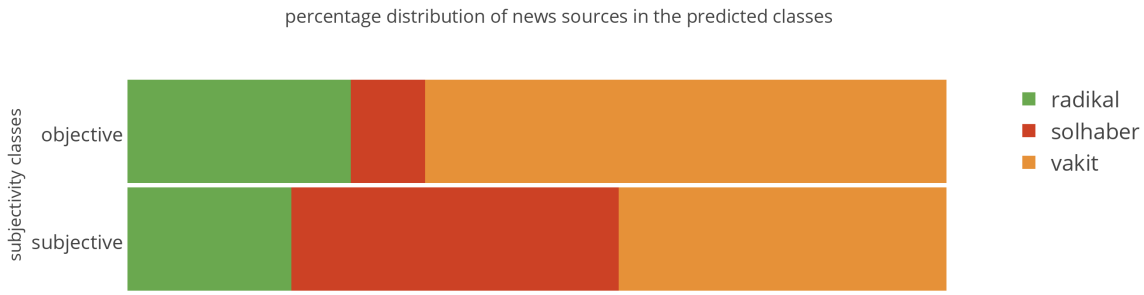

Figure 5.4. Percentage distribution of news sources in the predicted classes.

6. CONCLUSION & FUTURE WORK

Subjectivity analysis, as its name suggests, is one of the challenging tasks in computational linguistics. Overviewing the background work in the field, it appears that the analysis of subjectivity is relatively understudied compared to sentiment analysis. While the efforts on sentiment analysis in Turkish language appear to be increasing currently, there was none on subjectivity detection, to our knowledge. In this work, we developed a supervised subjectivity detection framework that runs at the document-level. A set of subjectivity-annotated news texts is created and prepared for further use. We applied a novel feature extraction mechanism to benefit more from the linguistic features to capture subjectivity and showed that it gave remarkably good results. In addition, we studied various dimensions of subjectivity classification that might influence the learning performance such as the intensity of subjectivity, dataset reliability and learning algorithms. Overall, we brought promising results along with new questions for further study.

As a future work, we would like to run our system on datasets from different domains and in different languages in order to examine the applicability of our features. Morphological properties of Turkish (or any other morphologically-rich language) can be practiced in more detail within the scope of feature extraction to tailor such a subjectivity detecting system for a more accurate but language-specific framework.

It is possible to make use of a list of emotion words to add a sentiment or emotion finding module to our subjectivity detector or to enhance the current subjectivity lexicons that we have. In [88, 89], around 150 Turkish words are listed with their values in three emotion dimensions, valence, arousal and dominance. Interpretation of these dimensions give clues about the polarity values of words. Indeed, in [88], a collection of positive, negative and neutral words is also presented based on emotion dimensions but the lists are small in terms of the number of words and the theoretical background to extract polarity values of those words needs to be strengthened. If we have a richer list of polarity words, we can calculate some subjectivity-related features

using it. For example, polar words can be supportive for enlarging the subjectivity-abstractness lexicon; bigrams and trigrams containing polar words can be counted as features for being possible idiomatic expressions; separate detection of subjectivity and sentiment(polarity) can be held so that we can investigate their relationship.

Additional analyses such as metaphor detection or objectivity clues collection, as a counterpart to subjectivity clues, can be devised. To increase accuracy, named entity recognition can be applied. Unsupervised learning methods, in particular soft clustering algorithms and topic models, can be applied to investigate the texts individually. Exploratory visualization approaches can help make the system more user-friendly for public use.

APPENDIX A: LEXICONS

A.1. List of Abstract Nouns

acı (*pain*)
affetme (*forgiveness*)
arzu (*desire*)
asalet (*nobility*)
ayıp (*ignominy*)
azim (*perseverance*)
bağlılık (*dependence*)
başarı (*success*)
başarısızlık (*failure*)
beğeni (*fondness*)
bela (*trouble*)
belirsizlik (*uncertainty*)
bencillik (*selfishness*)
bilinmezlik (*obscurity*)
bıkkınlık (*weariness*)
boşboğazlık (*indiscretion*)
buhran (*crisis*)
bunalım (*depression*)
bunaltı (*mope*)
büyü (*charm*)
canayakınlık (*amiability*)
çaresizlik (*helplessness*)
cefa (*torment*)
cehennem (*hell*)
çelişki (*contradiction*)
cennet (*heaven*)
cesaret (*courage*)
ciddiyet (*seriousness*)
çile (*ordeal*)
çirkinlik (*ugliness*)
çıldırma (*insanity*)
cömertlik (*generosity*)
coşku (*rhapsody*)
cüret (*daring*)
dalgınlık (*preoccupation*)
dayanıklılık (*endurance*)
dayanışma (*solidarity*)
dehşet (*horror*)
dert (*nuisance*)
dikkat (*attention*)
direnç (*resistance*)
doğruluk (*truth*)
dostluk (*friendship*)
dürüstlük (*honesty*)
düşsel (*imaginary*)
duyarlık (*sensitivity*)
duygu (*feeling*)
ebediyet (*eternity*)
elem (*woe*)
endişe (*anxiety*)
erdem (*virtue*)
fantezi (*fantasy*)
fazilet (*merit*)
fedakarlık (*sacrifice*)
felaket (*disaster*)
felsefe (*philosophy*)
gam (*gloom*)
gerginlik (*distemper*)
gizem (*mystery*)
gönül (*heart*)
güçlük (*difficulty*)
günah (*sin*)
gurbet (*homesickness*)
gurur (*pride*)
güven (*trust*)
güvensizlik (*distrust*)
güzellik (*beauty*)
hakaret (*revilement*)
haksızlık (*inequity*)
hasret (*longing*)
hassasiyet (*delicacy*)
hayal (*dream*)
hayalet (*ghost*)
haysiyet (*dignity*)
haz (*zest*)

heves (*enthusiasm*)
heyecan (*excitement*)
hiddet (*fury*)
his (*sense*)
hissiyat (*emotion*)
hırçınlık (*peevishness*)
hırs (*ambition*)
hoşgörü (*complaisance*)
hoşluk (*pleasantness*)
hoşnutsuzluk (*discontent*)
hürriyet (*liberty*)
hüsran (*frustration*)
hüzün (*sadness*)
huzur (*serenity*)
huzursuzluk (*unrest*)
içtenlik (*sincerity*)
ideal (*desire*)
ihanet (*infidelity*)
ihtiras (*ambition*)
ikiyüzlülük (*double-dealing*)
ilgi (*concern*)
iltifat (*compliment*)
iman (*creed*)
imkansızlık (*impossibility*)
inanç (*belief*)
inançsızlık (*impiousness*)
intikam (*revenge*)
insaniyet (*humaneness*)
insanlık (*humanity*)
irade (*will*)
itaat (*obedience*)
iyilik (*favour*)
ıstırap (*misery*)
ızdırap (*misery*)
kabus (*nightmare*)
kararlılık (*determination*)
kaygı (*disquiet*)
keder (*grief*)
kepazelik (*ignominy*)
keyif (*pleasure*)
kibarlık (*politeness*)
kibir (*arrogance*)
kin (*hatred*)
kırgınlık (*resentment*)
kıyamet (*doomsday*)
konfor (*comfort*)
korku (*fear*)
kötülük (*evil*)
kuşku (*suspicion*)
lezzet (*flavour*)
maneviyat (*spirituality*)
masumiyet (*innocence*)
melek (*angel*)
memnuniyet (*gladness*)
merhamet (*mercy*)
metanet (*resoluteness*)
methiye (*praise*)
mızmızlık (*crab*)
moral (*morale*)
motivasyon (*motivation*)
musibet (*calamity*)
mutluluk (*happiness*)
mutsuzluk (*unhappiness*)
namus (*purity*)
nankörlük (*ingratitude*)
nefret (*hate*)
neşe (*cheer*)
nezaket (*courtesy*)
niyet (*intention*)
öc (*revenge*)
olgunluk (*maturity*)
ölümsüzlük (*immortality*)
önyargı (*bias*)
övgü (*tribute*)
özgürlük (*freedom*)
özlem (*craving*)
özveri (*altruism*)
paranoya (*paranoia*)
peri (*fairy*)
rahatlık (*ease*)
rezalet (*disgrace*)
riya (*insincerity*)
ruh (*spirit*)
saadet (*felicity*)
sabır (*patience*)
sabırsızlık (*impatience*)

saçmalık (*nonsense*)
sadakat (*faithfullness*)
saflık (*naivety*)
samimiyet (*sincerity*)
şans (*luck*)
şaşkınlık (*confusion*)
saygı (*respect*)
şefkat (*compassion*)
şehvet (*lust*)
sempati (*sympathy*)
şer (*evil*)
sevap (*deed*)
sevgi (*affection*)
sevimsizlik (*bleakness*)
sevinç (*rejoicing*)
şevk (*incentive*)
şeytan (*devil*)
şiddet (*violence*)
sihir (*magic*)
sitem (*reproach*)
sıkıntı (*boredom*)
sıla (*homeland*)
sonsuz (*infinite*)
sonsuzluk (*infinity*)
sorun (*challenge*)
soyut (*abstract*)
sükunet (*tranquility*)
sükûnet (*tranquility*)
şükür (*glorification*)
şüphe (*doubt*)
sürpriz (*surprise*)
tasa (*worry*)
tedirginlik (*perturbation*)
tehlike (*danger*)
telaş (*haste*)
tereddüt (*hesitation*)
terör (*terror*)
tevazu (*humbleness*)
tılsım (*talisman*)
tuhaflık (*oddity*)
tutku (*passion*)
uğur (*talisman*)
uğursuzluk (*curse*)
ulaşılmazlık (*inaccessibility*)
ümit (*expectation*)
umut (*hope*)
umutsuzluk (*despair*)
utanç (*shame*)
ütopya (*utopia*)
üzüntü (*sorrow*)
vahşet (*atrocity*)
vefa (*fidelity*)
vicdan (*conscience*)
vuslat (*convergency*)
yalnızlık (*loneliness*)
yergi (*satire*)
yücelik (*glory*)
yürek (*heart*)
zeka (*intelligence*)
zevk (*enjoyment*)
zorluk (*hardship*)
zulüm (*cruelty*)

A.2. List of Subjective Verbs

affet (*forgive*)
ajite (*agitate*)
akla (*acquit*)
arlan (*ashamed*)
arzula (*desire*)
aydınlan (*enlighten*)
azarla (*scold*)
azmet (*persevere*)
azmettir (*enthuse*)
aşağıla (*humiliate*)
aşık (*love*)
baltala (*undermine*)
bastır (*supress*)
bağışla (*condone*)
becer (*accomplish*)
benzet (*liken*)
beğen (*acclaim*)
beğendir (*recommend*)
bloke (*block*)
bulan (*blur*)
bulandır (*blur*)
burul (*screw*)
büyüle (*charm*)
büyülen (*get-enthralled*)
canlandır (*invigorate*)
cesaretlen (*take-courage*)
cesaretlendir (*encourage*)
cezalandır (*punish*)
cezbet (*fascinate*)
coştur (*exhilarate*)
darıl (*take-offence*)
dayanış (*be-in-solidarity*)
destekle (*endorse*)
didikle (*tease*)
diret (*insist*)
domine (*dominate*)
donakal (*petrify*)
duygulan (*emphatize*)
duygulandır (*affect*)
düşkün (*fond*)
düşle (*dream*)
düşman (*enemy*)
emin (*ensure*)
endişelen (*worry*)
endişendir (*perturb*)
engelle (*hamper*)
espri (*joke*)
etkile (*influence*)
eğlen (*revel*)
eğlendir (*entertain*)
farkında (*aware*)
feda (*sacrifice*)
gizle (*dissemble*)
güldür (*amuse*)
güzelle (*flatter*)
gıcık (*irritate*)
gıpta (*admire*)
hakir (*despise*)
harap (*devastate*)
haykır (*exclaim*)
hayran (*enthuse*)
hayret (*astonish*)
heveslen (*aspire*)
heveslendir (*motivate*)
heyecanlan (*be-excited*)
heyecanlandır (*excite*)
hisset (*feel*)
hissettir (*evoke*)
hor (*insult*)
horgör (*insult*)
hoşgör (*tolerate*)
hücum (*assail*)
hürmet (*venerate*)
idealize (*idealize*)
ihbar (*denounce*)
ikna (*convince*)
ilgilen (*take-care-of*)
ilham (*inspire*)
imha (*annihilate*)
imren (*aspire*)
inan (*believe*)
inat (*persevere*)
isabet (*destine*)

istikrarsızlaştır (*destabilize*)
istirham (*request*)
itham (*allege*)
itibar (*accredit*)
itibarsızlaştır (*defame*)
itimat (*confide*)
iğren (*detest*)
işkillen (*suspect*)
karikatürize (*caricature*)
kasvetlen (*get-gloomed*)
kasvetlendir (*gloom*)
kaygılan (*worry*)
kederlen (*get-somber*)
kederlendir (*sadden*)
keyiflen (*rejoice*)
kirlet (*defile*)
kork (*fear*)
korkut (*fright*)
kov (*expel*)
kurtar (*liberate*)
kurtul (*recover*)
kutsa (*bless*)
kötüle (*denigrate*)
küs (*resent*)
küstür (*offend*)
küçümse (*disdain*)
kıllan (*mistrust*)
kırıl (*be-hurt*)
kıskan (*envy*)
kıskandır (*envy*)
kızdır (*rile*)
kızıştır (*provoke*)
kışkırt (*instigate*)
layık (*deserve*)
lütfet (*condescend*)
mahrum (*deprive*)
mahçup (*embarrass*)
manipüle (*manipulate*)
mağdur (*aggrieve*)
melun (*accurse*)
memnun (*delight*)
merak (*wonder*)
mest (*intoxicate*)
minnet (*appreciate*)
motive (*motivate*)
muhtaç (*impoverish*)
mutabık (*agree*)
mutenalaştır (*gentrify*)
müdafa (*plead*)
müdafaa (*plead*)
neşelen (*cheer*)
neşelendir (*cheer*)
niyetlen (*intent*)
onurlan (*be-honoured*)
onurlandır (*grace*)
onursuzlaştır (*dishonour*)
panikle (*panic*)
parla (*thrive*)
pislet (*besmirch*)
putlaş (*petrify*)
putlaştır (*idolize*)
pırılda (*shimmer*)
rahatla (*relax*)
rahatlat (*soothe*)
rahatsız (*distract*)
rencide (*offend*)
rezil (*disgrace*)
sabote (*sabotage*)
sabret (*endure*)
saldır (*attack*)
saldırt (*instigate*)
savun (*defend*)
serbest (*liberalize*)
sersemle (*dumbfound*)
sersemlet (*stupefy*)
sertleş (*harden*)
sertleştir (*harden*)
sev (*love*)
sevin (*cheer*)
sevindir (*elate*)
sez (*sense*)
sinirlen (*peeve*)
sinirlendir (*annoy*)
sivril (*excel*)

sokul (*ingratiate*)
soylulaş (*ennoble*)
soylulaştır (*gentrify*)
soysuzlaş (*get-degenerated*)
soysuzlaştır (*degenerate*)
suçla (*accuse*)
sıkıl (*be-bored*)
sızlan (*bemoan*)
tahakküm (*oppress*)
tahayyül (*imagine*)
tahrip (*devastate*)
takdir (*appreciate*)
tatmin (*satisfy*)
tavsiye (*suggest*)
tavır (*attitudinize*)
tebrik (*congratulate*)
tebriye (*exculpate*)
tedirgin (*disturb*)
tenezzül (*condescend*)
tenkit (*criticize*)
terbiye (*edify*)
terörize (*terrorize*)
teşekkür (*thank*)
teşvik (*embolden*)
tolere (*tolerate*)
tutum (*attitudinize*)
tölere (*tolerate*)
um (*anticipate*)
umursa (*mind*)
umutlan (*hope*)
umutsuzlaş (*despair*)
uslan (*sober*)
utan (*embarrass*)
utandır (*abash*)
uzlaş (*compromise*)
uzlaştır (*reconcile*)
yadsı (*deny*)
yaltaklan (*blandish*)
yaltaklık (*cringe*)
yalvar (*implore*)
yalvart (*implore*)
yanılt (*delude*)
yarat (*create*)
yazık (*spoil*)
yazıklan (*deplore*)
yağmala (*plunder*)
yen (*vanquish*)
yoksun (*deprive*)
yorul (*defatigate*)
yozlaş (*get-degenerated*)
yozlaştır (*degenerate*)
yücelt (*exalt*)
yükselt (*glorify*)
yıldır (*daunt*)
zafer (*triumph*)
zahmet (*toil*)
zarar (*damage*)
zorla (*compel*)
zorlan (*strain*)
zırvala (*twaddle*)
çabala (*endeavour*)
çarpıt (*contort*)
çatırda (*crepitate*)
çel (*beguile*)
çökert (*overthrow*)
çürüt (*corrupt*)
ödüllendir (*remunerate*)
öfkelen (*rage*)
öfkelendir (*exasperate*)
önesür (*claim*)
öv (*praise*)
övül (*be-praised*)
övün (*brag*)
özen (*elaborate*)
özgürleş (*liberalize*)
özgürleştir (*liberate*)
özle (*yearn*)
özlet (*yearn*)
ümitlen (*hope*)
ümitsizleş (*despair*)
ümitsizleştir (*dismay*)
ünlen (*repute*)
ürk (*scare*)
ürküt (*appall*)
üz (*afflict*)
ısrar (*insist*)
ışılda (*scintillate*)

şaka (*joke*)
şaşır (*buffle*)
şaşırt (*buffle*)
şenlen (*jollify*)
şenlendir (*jollify*)
şeref (*honour*)
şereflen (*be-honoured*)
şereflendir (*honour*)
şikayet (*complain*)
şok (*shock*)
şımar (*indulge*)
şımart (*indulge*)

APPENDIX B: ANNOTATION SYSTEM

In this part, we briefly present our online annotation system where 70 registered annotators labelled the news items, building the ground-truth set. After the user signs it, the web site[8] opens with a welcoming page, introduces the system like where the list of questions are available, indicates that the user can log out and come back to continue whenever she wants and asks the user to read our instructions for annotation.

Below we present the suggestions (in Turkish) that we introduced to the annotators about annotation.

Figure B.1. The first part of our 'Suggestions' page defining our objective and scope.

[8]Located at `http://engr.bilgi.edu.tr/dicle/`.

Figure B.2. The second part of our 'Suggestions' page describing our definition of objectivity in the scope this study along with a sample text per label.

Figure B.3. The third part of our 'Suggestions' page describing our definition of subjectivity in the scope this study along with a sample text per label.

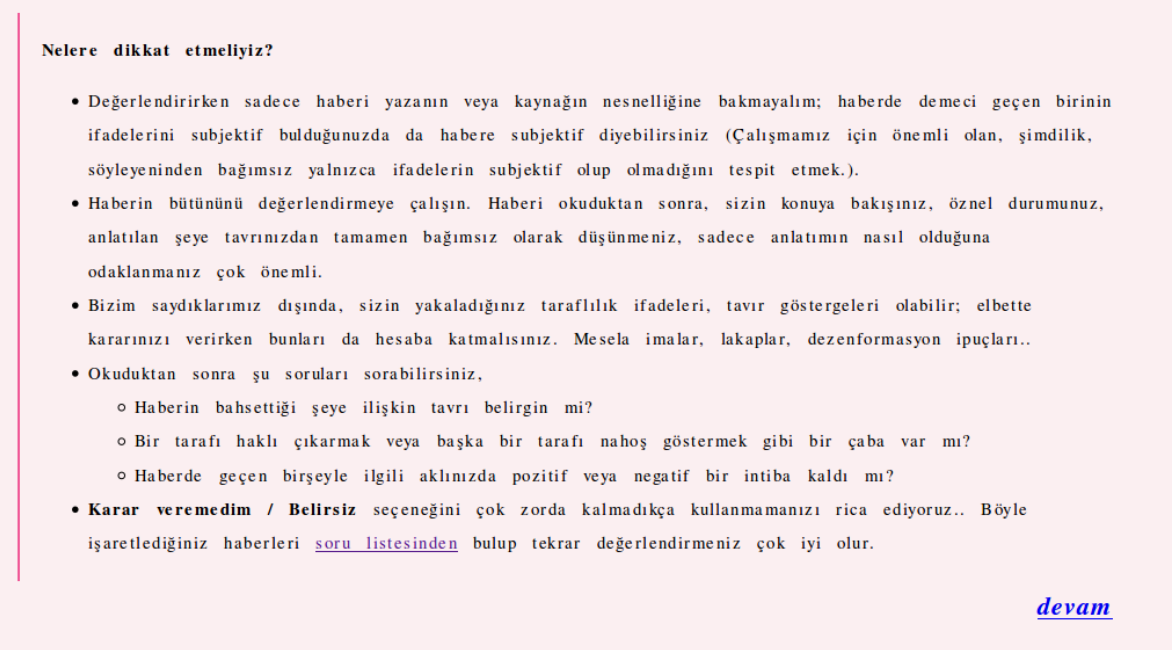

Figure B.4. The last part of our 'Suggestions' page detailing some remarks regarding the annotation process.

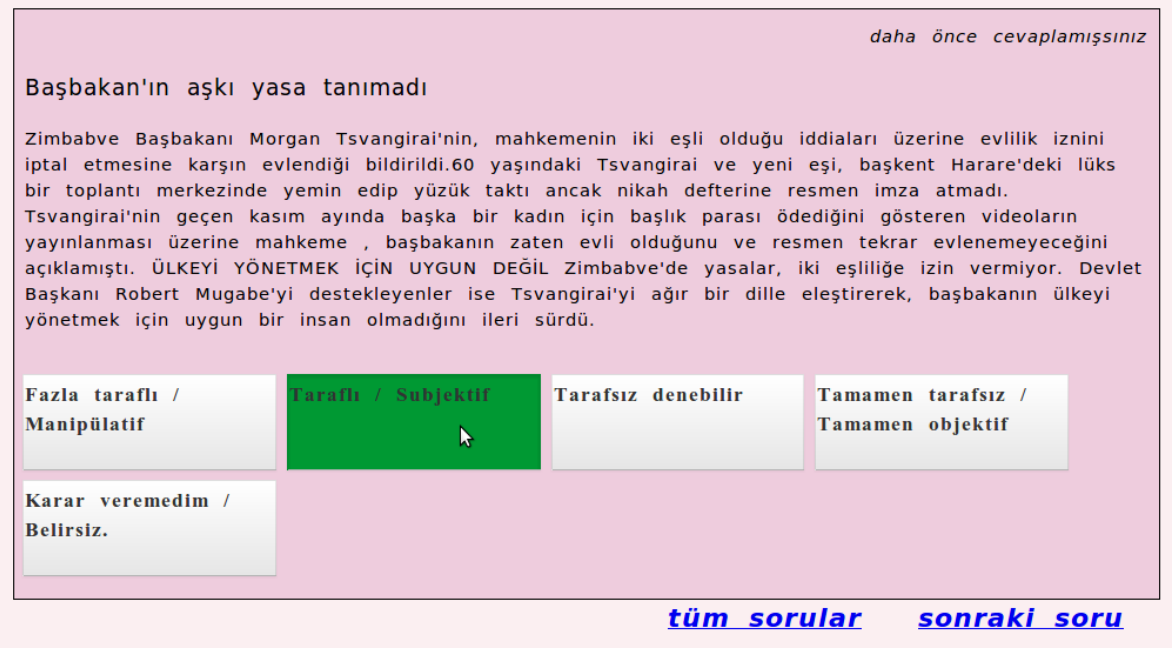

Figure B.5. An example annotation question.

REFERENCES

1. Pang, B. and L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1–135, 2008.

2. Esuli, A. and F. Sebastiani, "Determining Term Subjectivity and Term Orientation for Opinion Mining", *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, 2006.

3. Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts, "Learning Word Vectors for Sentiment Analysis", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 142–150, Stroudsburg, PA, USA, 2011.

4. Paltoglou, G. and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media", *ACM Transactions on Intelligent Systems and Technology*, Vol. 3, No. 4, pp. 66:1–66:19, 2012.

5. Pang, B. 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*, ACL '04, Stroudsburg, PA, USA, 2004.

6. Wang, D. and Y. Liu, "A Cross-corpus Study of Unsupervised Subjectivity Identification based on Calibrated EM", *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pp. 161–167, Stroudsburg, PA, USA, 2011.

7. Wiebe, J. and E. Riloff, "Finding Mutual Benefit between Subjectivity Analysis and Information Extraction", *IEEE Transactions on Affective Computing*, Vol. 2, No. 4, pp. 175–191, 2011.

8. Wiebe, J. M. and W. J. Rapaport, "A Computational Theory of Perspective and Reference in Narrative", *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, ACL '88, pp. 131–138, Stroudsburg, PA, USA, 1988.

9. Wiebe, J., "Tracking Point of View in Narrative", *Computational Linguistics*, Vol. 20, pp. 233–287, 1994.

10. Banfield, A., *Unspeakable Sentences: Narration and Representation in the Language of Fiction*, Routledge & Kegan Paul, Boston, USA, 1982.

11. Quirk, R., S. Greenbaum, G. Leech and J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman, London, 1985.

12. Neviarouskaya, A., H. Prendinger and M. Ishizuka, "Textual Affect Sensing for Sociable and Expressive Online Communication", *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, pp. 218–229, Berlin, Heidelberg, 2007.

13. Lin, W.-H., T. Wilson, J. Wiebe and A. Hauptmann, "Which Side Are You On?: Identifying Perspectives at the Document and Sentence Levels", *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pp. 109–116, Stroudsburg, PA, USA, 2006.

14. Thaul Lehrman, M., C. Ovesdotter Alm and R. A. Proano, "Detecting Distressed and Non-distressed Affect States in Short Forum Texts", *Proceedings of the Second Workshop on Language in Social Media*, pp. 9–18, Montréal, Canada, 2012.

15. Yu, H. and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences", *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pp. 129–136, Stroudsburg, PA, USA, 2003.

16. Balahur, A., R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot,

M. Halkia, B. Pouliquen and J. Belyaeva, "Sentiment Analysis in the News", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta, 2010.

17. Bracewell, D. B., J. Minato, F. Ren and S. Kuroiwa, "Determining the Emotion of News Articles", *Proceedings of the 2006 international conference on Intelligent computing: Part II*, ICIC'06, pp. 918–923, Berlin, Heidelberg, 2006.

18. Raaijmakers, S. and W. Kraaij, "A Shallow Approach to Subjectivity Classification", *Proceedings of the International Conference on Weblogs and SocialMedia (ICWSM)*, 2008.

19. Kaya, M., G. Fidan and I. H. Toroslu, "Sentiment Analysis of Turkish Political News", *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pp. 174–180, Washington, DC, USA, 2012.

20. Fukuhara, T., H. Nakagawa and T. Nishida, "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events", *Proceedings of the International Conference on Weblogs and SocialMedia (ICWSM)*, 2007.

21. Çetin, M. and M. F. Amasyalı, "Active Learning for Turkish Sentiment Analysis", *Innovations in Intelligent Systems and Applications (INISTA)*, 2013.

22. Melville, P., W. Gryc and R. D. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 1275–1284, New York, NY, USA, 2009.

23. Scheible, C. and H. Schütze, "Unsupervised Sentiment Analysis with a Simple and Fast Bayesian Model Using Part-of-Speech Feature Selection", J. Jancsary (Editor), *Proceedings of KONVENS 2012*, pp. 269–273, 2012.

24. Lin, C., Y. He and R. Everson, "A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection", *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pp. 144–152, Stroudsburg, PA, USA, 2010.

25. Turney, P. D., "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424, Stroudsburg, PA, USA, 2002.

26. Demirtaş, E. and M. Pechenizkiy, "Cross-lingual Polarity Detection with Machine Translation", *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pp. 9:1–9:8, New York, NY, USA, 2013.

27. Blitzer, J., M. Dredze and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 440–447, Prague, Czech Republic, 2007.

28. Li, F., S. J. Pan, O. Jin, Q. Yang and X. Zhu, "Cross-domain Co-extraction of Sentiment and Topic Lexicons", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pp. 410–419, Stroudsburg, PA, USA, 2012.

29. Korayem, M., D. J. Crandall and M. Abdul-Mageed, "Subjectivity and Sentiment Analysis of Arabic: A Survey", *Advanced Machine Learning Technologies and Applications - First International Conference, AMLTA 2012, Cairo, Egypt, December 8-10, 2012. Proceedings*, Vol. 322 of *Communications in Computer and Information Science*, pp. 128–139, 2012.

30. Veselovská, K., jr. Jan Hajič and J. Šindlerová, "Creating Annotated Resources for Polarity Classification in Czech", *Empirical Methods in Natural Language Pro-*

cessing - Proceedings of the Conference on Natural Language Processing 2012, pp. 296–304, Wien, Austria, 2012.

31. Mihalcea, R., C. Banea and J. Wiebe, "Learning Multilingual Subjective Language via Cross-Lingual Projections", *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 976–983, Association for Computational Linguistics, Prague, Czech Republic, June 2007.

32. Nakagawa, T., K. Inui and S. Kurohashi, "Dependency Tree-based Sentiment Classification Using CRFs with Hidden Variables", *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 786–794, Stroudsburg, PA, USA, 2010.

33. Banea, C., R. Mihalcea, J. Wiebe and S. Hassan, "Multilingual Subjectivity Analysis Using Machine Translation", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 127–135, Association for Computational Linguistics, Honolulu, Hawaii, 2008.

34. Wan, X., "Co-Training for Cross-Lingual Sentiment Classification", *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 235–243, Association for Computational Linguistics, Suntec, Singapore, August 2009.

35. Meng, X., F. Wei, X. Liu, M. Zhou, G. Xu and H. Wang, "Cross-Lingual Mixture Model for Sentiment Classification", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 572–581, Association for Computational Linguistics, Jeju Island, Korea, 2012.

36. Scholz, T., S. Conrad and L. Hillekamps, "Opinion Mining on a German Corpus of a Media Response Analysis", P. Sojka, A. Horák, I. Kopeček and K. Pala (Editors), *Text, Speech and Dialogue*, Vol. 7499 of *Lecture Notes in Computer Science*, pp. 39–46, Springer Berlin Heidelberg, 2012.

37. Ng, V., S. Dasgupta and S. M. N. Arifin, "Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews", *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions*, pp. 611–618, 2006.

38. Toprak, C. and I. Gurevych, "Document Level Subjectivity Classification Experiments in DEFT'09 Challenge", *Proceedings of the DEFT'09 Text Mining Challenge*, pp. 89–97, Paris, France, 2009.

39. Wiebe, J., T. Wilson, R. Bruce, M. Bell and M. Martin, "Learning Subjective Language", *Computational Linguistics*, Vol. 30, No. 3, pp. 277–308, 2004.

40. Wilson, T., J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-level Sentiment Analysis", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Stroudsburg, PA, USA, 2005.

41. Nasukawa, T. and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing", *Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, pp. 70–77, ACM, New York, NY, USA, 2003.

42. Dasgupta, S. and V. Ng, "Topic-wise, Sentiment-wise, or Otherwise?: Identifying the Hidden Dimension for Unsupervised Text Classification", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pp. 580–589, Stroudsburg, PA, USA, 2009.

43. Wang, S. and C. D. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pp. 90–94, Stroudsburg, PA, USA, 2012.

44. Pang, B., L. Lee and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques", *Proceedings of the ACL-02 Conference on*

Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, pp. 79–86, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002.

45. Riloff, E., J. Wiebe and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping", *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pp. 25–32, Edmonton, Canada, 2003.

46. Miller, G. A., "WordNet: A Lexical Database for English", *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

47. Rennison, E., "Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes", *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, UIST '94, pp. 3–12, ACM, New York, NY, USA, 1994.

48. Zhao, J., F. Chevalier, C. Collins and R. Balakrishnan, "Facilitating Discourse Analysis with Interactive Visualization", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp. 2639–2648, 2012.

49. Oliver, S., G. Gali, F. Chevalier and S. Diamond, "Discursive Navigation of Online News", *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pp. 82–85, ACM, New York, NY, USA, 2012.

50. Boynukalın, Z., *Emotion Analysis of Turkish Texts by Using Machine Learning Methods*, M.S. Thesis, Middle East Technical University, 2012.

51. Akkaya, C., J. Wiebe, A. Conrad and R. Mihalcea, "Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis", *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 87–96, Association for Computational Linguistics, Portland, Oregon, USA, 2011.

52. Pennebaker, J. W., M. E. Francis and R. J. Booth, *Linguistic Inquiry and Word Count*, Lawerence Erlbaum Associates, Mahwah, NJ, 2001.

53. Stone, P. J., D. C. Dunphy, M. S. Smith and D. M. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, 1966.

54. Bradley, M. M. and P. J. Lang, *Affective Norms for English Words (ANEW): Stimuli, Instruction Manual, and Affective Ratings*, Tech. rep., Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

55. Baccianella, S., A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta, 2010.

56. Ruppenhofer, J., M. Ellsworth, M. R. Petruck, C. R. Johnson and J. Scheffczyk, *FrameNet II: Extended Theory and Practice*, International Computer Science Institute, Berkeley, California, 2006.

57. Osherenko, A. and E. André, "Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect?", *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, ACII '07, pp. 230–241, Berlin, Heidelberg, 2007.

58. Pennebaker, J., M. Mehl and K. Niederhoffer, "Psychological Aspects of Natural Language Use: Our Words, Our Selves", *Annual Review of Psychology*, Vol. 54, No. 1, pp. 547–577, 2003.

59. Eroğul, U., *Sentiment Analysis in Turkish*, M.S. Thesis, Middle East Technical University, 2009.

60. Scherer Klaus R., H. G., Wallbott, "Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning", *Journal of Personality and Social Psychology*, Vol. 66(2), pp. 310–328, 1994.

61. Vural, A. G., B. B. Cambazoğlu and P. Şenkul, "Sentiment-focused Web Crawl-

ing", *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pp. 2020–2024, ACM, New York, NY, USA, 2012.

62. Vural, A. G., *Sentiment-focused Web Crawling*, Ph.D. Thesis, The Graduate School of Natural and Applied Sciences of Middle East Technical University, Ankara, Turkey, 2013.

63. Vural, A. G., B. B. Cambazoğlu, P. Şenkul and Z. Özge Tokgöz, "A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish", *Computer and Information Sciences III*, pp. 437–445, 2012.

64. Thelwall, M., K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment in Short Strength Detection Informal Text", *Journal of the Association for Information Science and Technology*, Vol. 61, No. 12, pp. 2544–2558, 2010.

65. Liu, B., "Sentiment Analysis and Subjectivity", N. Indurkhya and F. J. Damerau (Editors), *Handbook of Natural Language Processing, Second Edition*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

66. Richardson, J. E., *Analysing Newspapers: An Approach from Critical Discourse Analysis*, Palgrave Macmillan, New York, NY, 2007.

67. Wittgenstein, L., *Philosophical Investigations*, Oxford : Blackwell, 3rd edn., 2001.

68. Walton, D., *Fundamentals of Critical Argumentation*, Cambridge University Press, New York, NY, USA, 2006.

69. Johnson, R. H. and J. A. Blair, *Logical Self-Defense*, McGraw-Hill Ryerson, Toronto, 1983.

70. Potts, C., "Presupposition and Implicature", S. Lappin and C. Fox (Editors), *The Handbook of Contemporary Semantic Theory*, Wiley-Blackwell, 2 edn., To appear.

71. Stevenson, C. L., "Persuasive Definitions", *Mind*, Vol. 47, pp. 331–350, 1938.

72. Riloff, E. and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions", *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pp. 105–112, Stroudsburg, PA, USA, 2003.

73. Wiebe, J., T. Wilson and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language", *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210, 2005.

74. Cohen, J., "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement*, Vol. 20, No. 1, p. 37, 1960.

75. Shin, H., M. Kim, H. Jang and A. Cattle, "Annotation Scheme for Constructing Sentiment Corpus in Korean", *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pp. 181–190, Faculty of Computer Science, Universitas Indonesia, Bali,Indonesia, 2012.

76. Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*, Sage, 2nd edition edn., 2004.

77. Artstein, R. and M. Poesio, "Inter-coder Agreement for Computational Linguistics", *Computational Linguistics*, Vol. 34, No. 4, pp. 555–596, 2008.

78. Hatzivassiloglou, V. and K. McKeown, "Predicting the Semantic Orientation of Adjectives", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pp. 174–181, Madrid, Spain, 1997.

79. Wiebe, J., "Learning Subjective Adjectives from Corpora", *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pp. 735–740, 2000.

80. Sak, H., T. Güngör and M. Saraçlar, "Turkish Language Resources: Morphological

Parser, Morphological Disambiguator and Web Corpus", *Proceedings of the 6th International Conference on Advances in Natural Language Processing*, GoTAL '08, pp. 417–427, Springer-Verlag, Berlin, Heidelberg, 2008.

81. Tekcan, A. I. and I. Göz, *Türkçe Kelime Normları: 600 Türkçe Kelimenin İmgelem, Somutluk, Sıklık Değerleri ve Çağrışım Setleri*, Boğaziçi Üniversitesi, İstanbul, 2005.

82. Cortes, C. and V. Vapnik, "Support-Vector Networks", *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

83. Manning, C. D., P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

84. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

85. Wu, K.-P. and S.-D. Wang, "Choosing the Kernel Parameters for Support Vector Machines by the Inter-cluster Distance in the Feature Space", *Pattern Recognition*, Vol. 42, No. 5, pp. 710–717, 2009.

86. Turney, P. D. and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research*, Vol. 37, No. 1, pp. 141–188, 2010.

87. Dias, G., D. D. Lambov and V. Noncheva, "High-level Features for Learning Subjective Language across Domains", *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM*, California, USA, 2009.

88. Kılıç, A., *Age related Changes in Recognition Memory for Emotional Stimuli*, M.S.

Thesis, Middle East Technical University, 2007.

89. Başgöze, Z., *Emotional Conflict Resolution in Healthy and Depressed Populations*, M.S. Thesis, Middle East Technical University, 2008.