İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

Cumali TÜRKMENOĞLU

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

OCAK 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERDE DUYGU ANALİZİ

YÜKSEK LİSANS TEZİ

Cumali TÜRKMENOĞLU
(504111541)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ

OCAK 2015

İTÜ, Fen Bilimleri Enstitüsü'nün 504111541 numaralı Yüksek Lisans Öğrencisi **Cumali TÜRKMENOĞLU**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"TÜRKÇE METİNLERDE DUYGU ANALİZİ"** başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Yrd. Doç. Dr. Gülşen CEBİROĞLU ERYİĞİT** İstanbul Teknik Üniversitesi	.....
	Doç. Dr. Banu DİRİ Yıldız Teknik Üniversitesi	.....
		.....

Teslim Tarihi : **13 Aralık 2014**
Savunma Tarihi : **23 Ocak 2015**

Aileme,

ÖNSÖZ

Çalışmalarım sırasında bilgi ve tecrübelerini benden esirgemeyen değerli hocam ve tez danışmanım Sayın Yrd. Doc. Dr. Ahmet Cüneyd TANTUĞ'a ve tez yazım sürecinde bana destek olan sevgili eşime teşekkürlerimi sunarım.

Ocak 2015

Cumali TÜRKMENOĞLU
(Araştırma Görevlisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

BÇ : Bilgi Çıkarımı
BK : Bilgi Kazancı
BKO : Bilgi Kazancı Oranı
DA : Duygu Analizi
DDİ : Doğal Dil İşleme
DF : Döküman Frekansı (Document Frequency)
DÖ : Derin Öğrenme
KA : Karar Ağaçları
KDM : Karar Destek Makineleri
MÖ : Makine Öğrenmesi
NB : Naive Bayes
POS : Part Of Speech
SDM : Sonlu Durum Makineleri
TF : Terim Frekansı (Term Frequency)

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

TÜRKÇE METİNLERDE DUYGU ANALİZİ

ÖZET

Başkalarının ne düşündüğü, biz insanlar için her zaman merak konusu olmuştur. "İnsanlar ne düşünüyor?" sorusu, aynı zamanda üretim, pazarlama, hizmet ve reklamcılık firmaları için de toplumun, ürün, hizmet ve marka isimleri hakkındaki görüşlerini öğrenmeleri açısından son derece önemlidir. Genel olarak firmalar, kullanıcı/müşteri analizini, ya müşterilerden görüşlerini içeren geri bildirim formları toplayıp, elle analiz edip, çıkarımlarda bulunarak ya da bir anket firmasına yüklü miktarlar ödeyerek, anketler ile yapmaya çalışırlar. Ancak bu yöntemler istatistiksel olarak geniş kitlelere ulaşılmasında çok yetersiz, insan emeği bağlamında masraflı yöntemlerdir. Sosyal medya platformları ve diğer internet ortamları özellikle hedef kitleden geri dönüşüm alabilmek için önemli ve yeterince geniş kaynaklardır. Ancak bunları insan eliyle analiz etmek neredeyse imkansızdır. Bu noktada Duygu (Sentiment) Analizi (DA) araçları devreye girerler ki bunların, sosyal platformları gözlemlemek için en işlevsel araçlar olduğu söylenebilir.

Bir metinde ilgili konu hakkındaki tutum, ancak DA yapılarak anlaşılabilir. DA bir metnin duygu barındırıp barındırmadığı ve bu duygunun olumlu ya da olumsuz olma durumunun saptanması sürecidir [1]. Duygu barındıran metinler genellikle görüş ya da değerlendirme içerirler. Bu görüş ve analizler herhangi bir konu, şahıs, marka ya da siyasi görüş hakkında olabilir.

İnternet olmadan önce dijital ortamdaki veri ve bu veriye ulaşım imkanları çok kısıtlıydı. İnternet'in yaygınlaşmasıyla beraber insanlar belli konular hakkındaki görüşlerini forumlar, bloglar ve sosyal medya platformlarında paylaşmaya başladılar. Bu paylaşımlardaki görüşler, sosyal analiz ve anketler için olduğu gibi firmaların da kendileri, ürünleri veya hizmetleri ile ilgili araştırmaları ve analizleri için değerli kaynaklar oluştururlar. Dijital platformlarda hızla biriken büyük oranda verinin insanlar tarafından işlenmesi çok zor olduğundan otomatik olarak bilgisayarlarla yapılması kaçınılmaz bir durum olmuştur. Bu verilerin bilgisayarlar tarafından hızlı bir şekilde işlenebilmesi ise, bu kaynakların, piyasada kullanılabilmesine imkan sağlamıştır.

Duygu analizi, Doğal Dil İşleme (DDİ) ve metin madenciliği için zor bir çalışma alanıdır. Piyasa değerinin olması ve pratik sonuçlar alınabilmesi hem akademik çalışmaların hem de endüstrinin bu alana ciddi bir şekilde yönelmesini sağlamıştır. Ancak görüş/duygu bildiren kaynakları WEB üzerinde bulmak ve onlara ulaşıp işlemek hâlâ zorlu bir görev olarak karşımızda durmaktadır. Çünkü her biri büyük miktarda görüş/duygu barındıran geniş sayıda farklı kaynaklar mevcuttur ve bu kaynakların birçoğunda, görüş/duygu uzun metinler içerisinde gizli bir şekilde yer alır. Bir insan için ilgili kaynakları bulmak, o kaynaklardan ilgili görüş/duygu içeren kısımları bulup onları özetlemek ve kullanılabilir bir biçimde organize etmek çok zor

ve zahmetli bir iştir. Bundan dolayı, otomatik olarak görüş/duygu keşfetmek, analiz etmek ve özetlemek için özel sistemlere ihtiyaç vardır. Görüş madenciliği olarak da bilinen DA, bu ihtiyaçlardan doğar.

Sosyal analiz ve anketlerin otomatik olarak yapılabilmesi için DA'nin büyük veri kümeleri üzerinde otomatik olarak bilgisayarlara yaptırılması gerekir. Otomatik DA yapılırken belli teknikler kullanılmaktadır. Bunlardan en çok kullanılanları Makine Öğrenmesi (MÖ) ve sözlük tabanlı DA tekni̇kleridir. Bu tekniklerin kullanıldığı çalışmaların birçoğu İngilizce üzerinde yoğunlaşmasına rağmen diğer diller için de çalışmalar popüler olmaya başlamıştır.

Bu tez çalışmasında hem İngilizce hem de Türkçe için yapılan çalışmalarda kullanılan MÖ ve sözlük tabanlı DA metotları yeni özellikler eklenerek oluşturulup farklı iki veri kümesi üzerinde değerlendirildi.

Bu çalışmada daha önce İngilizce ve Türkçe için çalışılmış metotlardan MÖ ve sözlük tabanlı DA olmak üzere iki ayrı DA metodu Türkçe için gerçeklenmiştir. Bu metotları kısa ve uzun metinler olmak üzere iki farklı Türkçe veri kümesine uygulayıp başarımlarını ölçtük. Türkçenin yapısal özelliklerini de göz önüne alacak şekilde ön işlemler uyguladı. Öncelikle bir deasciifying (Türkçeleştirme) ve düşük seviye normalleştirme uygulanarak Türkçeye uygun yazılmayan ve çok tekrarlı harfler içeren kelimeler düzeltildi. Kelimeler asıl anlamlarını köklerinde barındırdığıdan gereksiz ekler atılıp, asıl anlamı içeren kelime kökülerine ulaşıldı. Bunu yaparken varlık/yokluk (-lı,-li,-sız,-siz) eklerini ve olumsuzluk bildiren ekleri (-me,-ma) ya kaldırmadık ya da ona özel bir işaret koyarak muhafaza ettik. Şimdiye kadar yapılan çalışmalardan farklı olarak hem MÖ hem de sözlük tabanlı DA için bazı yeni özellikler eklendi. Bu yeni özellikler sözlük tabanlı DA için bileşik kelimeler ve varlık/yokluk eki barındıran kelimelerdir.

Sözlük tabanlı DA için her kelimesi taşıdığı duygu yönelimine göre puanlandırılmış bir sözlük kullanılarak bir metnin duygu yönelimini bulmaya çalıştık. Kullanılan sözlüğü oluşturmak için Thellwal ve diğ. [2] çalışmalarında kullanıldıkları Sentistrength sözlüğünü Türkçeye çevrildi. Ayrıca sözlükte eksik olan diğer gerekli sözcükler, birleşik kelimeler ve varlık/yokluk eki barındıran kelimeler eklemek kaydıyla sözlük genişlettilmiştir. Kelime köklerine inildiğinden dolayı kelimler ad-sıfat veya fill olmak üzere iki ayrı etiketle etiketlenmiştir. Sözlük tabanlı DA olumlu-olumsuz ve olumlu-olumsuz-nötr senaryoları olarak uygulandı ve sonuçları değerlendirildi. Genel olarak, işlenecek metindeki kelimelerden duygu sözlüğünde yer alanlarının sözlükteki puanlarının toplanmasıyla elde edilen puana göre sınıflandırma yapılmıştır. Diğer taraftan MÖ tabanlı DA için n-gram'lar öznitelik olarak bag-of-words şeklinde kullanılmıştır.

Bu iki metodun güçlü ve zayıf yönlerini görülebilmesi için kısa ve uzun yorum metinler içeren iki farklı veri kümesi üzerinde test edildi. Bu veri kümeleri; diğerine kıyasla daha kısa ve kuralsız yorumlardan oluşan Twitter veri kümesi ve görece daha uzun, nispeten daha kurallı yazılmış film yorumlarından oluşan Film Yorumları veri kümesidir. Twitter veri kümesine uygulandığında sözlük tabanlı DA metodu ile %75.2, MÖ tabanlı DA metodu ile ise, Karar Destek Makineleri (KDM) sınıflandırıcısı kullanılarak, %85 başarı elde edilmiştir. Film yorumları veri kümesine uygulandığında ise sözlük tabanlı DA metodu ile %79,5, MÖ tabanlı DA metodu ile KDM sınıflandırıcısı kullanılarak %89 başarı elde edilmiştir. Twitter verisi, gramer ve yapısal kural eksikliğinden dolayı DDİ çalışmaları için ez zor verilerden biridir. Film

Yorumları veri kümesi daha kurallı ve düzgün metinlerden oluştuğundan ve tek hedefe (film) odaklı olduğundan, her iki yaklaşımda da daha iyi sonuç vermiştir.

Bu çalışmada birçok ön işlem uygulanmıştır. Bu ön işlemler duygu analizi ve özellikle Türkçe için önemlidirler. Bileşik kelimelerin yakalanması ve varlık/yokluk eklerinin kullanılması sözlük tabanlı duygu analizi yaklaşımında önemli etki yaratmıştır. Bu yöntemlerin etkisi, gizli bilgilerin ortaya çıkarılıp işlenmesinin umut verici olduğu göstermiştir. Gizli bilginin yanında varolan bilginin cümledeki hangi nesneye yönelik olduğu da çok önemlidir. Daha ileriki çalışmalar için bağlılık analizi yapılarak sadece ilgilendiğimiz nesne ile ilgili kelimelerin dikkate alınması sağlanabilir.

MÖ metodu KDM sınıflandırıcısıyla beraber, birçok çalışmada olduğu gibi bizim çalışmamızda da en yüksek başarımı sağlamaktadır. Fakat, eğitilme ihtiyacı olduğundan ve eğitim kümesi için büyük miktarlarda etiketli veri gerektiğinden, MÖ tabanlı yaklaşım tercih edilmeyebilir. Sözlük tabanlı duygu analizi için herhangi bir eğitim kümesi gerekmediğinden etiketleme işine de gerek kalmamaktadır. Oluşturduğumuz sözlük her ne kadar elle oluşturulmuş olsa da genel amaçlı olduğundan, dışarıdan girilen herhangi bir metni sözlüğe göre değerlendirip sınıflandırabileceğinden, alan bağımsızdır ve hiçbir eğitim kümesi gerektirmez.

Elimizdeki eğitim kümesini kullanarak daha alana özel bir sözlük oluşturmak ve başarımı daha da yükseltmek mümkündür. Ancak bu durumda sözlük tabanlı DA'nin avantajlarından olan eğitim seti gerektirmemeyi kaybetmiş oluruz. Sözlük tabanlı duygu analizi denetimsiz ve alan bağımsız bir çalışma olmasına rağmen elde ettiğimiz başarı umut vericidir.

SENTIMENT ANAYSIS IN TURKISH TEXT

SUMMARY

There is a remarkable curiosity inside us to know what others think. It is also important for production, marketing, service and advertising firms to learn the attitude of people towards their goods, brands and services. Firms used to monitor customer attitudes by receiving feedback forms from their customers and analyze them manually or made some questionnaire to survey companies with charge of money. However these methods were not able to capture statistically sufficient size masses and were costly in terms of human labor and money. Social media platforms, which are easily accessible platforms, provide remarkable sources to get feedbacks from target masses, but it is impossible to analyze these feedbacks by human labor. Therefore, automated sentiment analysis tools are crucial for companies' customer services to have the capability of capturing complaints and/or positive feedbacks in the right time. Processing by computers allows these data to be used in the market. Implementing an efficient sentiment analysis tool will increase the customer satisfaction and will decrease the costs. This is the motivation of sentiment analysis research area. We can say that: sentiment analysis is one of the most useful tool for social media monitoring.

A text with sentiment generally includes opinions, attitudes and evaluating. Opinions and attitudes can be towards a topic, a person, a brand or a politic view. They are not only valuable sources for social researches and surveys but also quite important for firms to analyze responses and feedbacks about their goods and services. Sentiment analysis is needed to capture the attitude of a text towards any topic.

Sentiment analysis is the process of determining whether a text includes sentiment or not and classifying the sentiment into positive, negative and neutral classes.

Although, sentiment analysis is a hard task for NLP and Data Mining research areas, giving practical solutions and having high market value results increasing academic and market interest in it. Accessing data including sentiment on WEB and processing it are still hard tasks to be solved, because there are huge number of sources including sentiment and most of sentiments are hidden in long texts.

Before development of WEB, there were almost no information on digital platforms and no possibility of access to this information. People started to share their opinions on certain topics on digital platforms. The amount of accessible information with opinion on the Web has been increasing with the contribution of forums, columns, blogs, and social media. Processing this information, extracting the subjectivity and classifying the sentiment are the main challenges of the sentiment analysis that need to be solved. Sarcasm and irony also have remarkable importance and interest in both psychology [3] and NLP [4] [5] research area. Increasing the accuracy in detecting sarcasm will increase the performance of the sentiment analysis. Unfortunately it is also a difficult task to identify the sarcasm in a natural text even for a human [4].

Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text [1]. Extracting opinions and analyzing the polarity of these opinions are the main problems of the sentiment analysis. Various approaches are utilized to solve these problems in academic researches. Most of them are on subjectivity classification and sentiment classification. Subjectivity classification is a problem of classifying any document as objective or subjective and sentiment classification is the classification of these subjective documents into positive or negative [1] classes according to their sentimental polarity. NLP and machine learning techniques are extensively used for Sentiment Analysis. Knowing the characteristics of the language are essential for NLP and Sentiment Analysis because different languages require different preprocessing techniques.

Sentiment analysis approaches are mainly based on either machine learning or lexicon based methods. Both methods have advantages and disadvantages in terms of accuracy and human labor. Our goal is to show the comparison of the strengths and weaknesses of these methods on two different types of datasets. As a lexicon based method, we build a framework similar to the systems described in Thelwall et al. [2] and Vural et al. [6]. To implement machine learning based sentiment analysis, we have investigated several machine learning methods like Pang et al. [7] and Eroğul [8].

The majority of sentiment analysis approaches are concentrated on English. However,there exists a number of sentiment analysis studies on Turkish [8] [6]. Eroğul [8] handled the sentimental analysis problem as a supervised machine learning classification problem and applied different ML techniques with different features like unigrams, bigrams, POS tags and combination of them. Vural et al. [6] presented a lexicon based sentiment analysis framework using Turkish version Sentistrength [2] lexicon. They used an approach based on summing lexicon scores of sentiment oriented words in related text. In this work, we applied both ML based and Lexicon based SA methods on Turkish with additional features.

In order to evaluate the performance of lexicon based and ML based sentiment analyzers, we use two datasets exhibiting different characteristics. Our first dataset is comprised of tweets which suffer from orthographic and grammatical problems. Tweets are usually difficult to process for NLP purposes since they frequently contain abbreviations, missing vocals that need devocalization and ungrammatical constructs both due to the character limitation of Twitter and mobile devices with limited text entry capabilities. We collect another dataset that consists of movie reviews which are more grammatical and orthographic than tweets. We applied our tests for binary (positive and negative) and trinary (positive-negative and neutral) classification. Pre-processing is one of the most important steps of the sentimental analysis in Turkish. Having a very productive inflectional and observational morphology, Turkish is a difficult language to process.

A number of preprocessing steps are required for both lexicon based and ML based approaches due to the productive Turkish morphology. In this study, we employ deasciification, basic text normalization, morphological analysis, morphological disambiguation and multi-words expressions extraction preprocessing steps. Text normalization pre-processing steps such as spelling correction are necessary prior to morphological analysis step since the data is noisy. A finite-state-machine based morphological analyzer [9] is used to produce root of words, suffixes and morphological tags. This level produce ambiguous results. Since the morphological

analysis stage produces ambiguous results, a morphological disambiguation module is required. We used a rule based morphological disambiguator [10]. Multi-words expressions extraction aims to identify the segments of the texts which are generally sequential but not compositional [11]. We use Kemal Oflazer's MWEs extraction application's Perl script to handle the MWEs extraction problem. Finally we identify and combine expressions which have different meanings and may have/haven't sentiment when they separate from each other, e.g. "kafayı ye-" (literally eat the head) none of the words have an sentiment polarity by their self but it means "to get mentally deranged" and has negative sentiment polarity when they are together. We added these sentiment holding MWEs to our lexicon.

Our lexicon based sentiment analysis approach depends on comparing features of a given text with a pre-determined sentimentally oriented lexicon. Sentiment analysis does not require a detailed pre-processing [12] phase before classification for English but it is necessary for Turkish and similar agglutinating languages. Turkish is an agglutinating language in which it is possible to add many suffixes to word roots. These derivational and inflectional suffixes can change the POS tag and sentimentally orientation of the word. Important suffixes for sentiment analysis are considered to be the negation suffix (**+ma/+me**) and absence/presence suffixes (**+sız/+siz (without), +lı/+li (with)**) which can change the sentiment orientation of a nominal word. Handling these suffixes increase the performance of the sentiment analysis [13] [7]. The morphological analysis is needed to handle linguistic features for sentiment analysis, e.g. roots, POS tags, suffixes and adjuncts of the words.

For a lexicon based sentiment analyzer, it is necessary to have a sentimentally oriented lexicon which is effective to detect the sentiment of a sentence. Since there were no Turkish lexicon we manually translated a basic English lexicon (Sentistrength, 2547 words) [2] into Turkish. Although there were some other more detailed lexicon in literature, such as SenticNet [14], WordNet-Affect [15], we used Sentistrength lexicon as a baseline lexicon. We reconstructed it by adding 700 MWEs, 650 words with absence/presence suffixes, 110 extra needed words for Turkish (slangs, curses and some special words) and we removed 350 root words due to adding them again as words with absence/presence suffixes. Actually we use Sentistrength as a starting point. After reconstructing, our final lexicon contains 2784 nominals and 873 verbs totally 3657 terms which have a polarity magnitude between [-5, +5].

Because of negation (-me, -ma) and absence/presence suffixes (**+sız/+siz (without), +lı/+li (with)**) suffixes, we must be careful when finding root of the words. It is not effective technique to use regular expressions like 'isolat*' which stands for 'isolate' 'isolated' 'isolation' 'isolating' in English, because of differentiation of words with suffixes in Turkish. Negation occurs in two different ways for Turkish. The first is using negation words ("değil", "yok") and second is using negation suffixes (–me, -ma).

When negation suffixes met we add negation word ("değil") after related words, so that all negation forms become standardized. During calculating the sentiment score of texts, negation words change the sign of the sentiment score of the related word.

We use a booster words list ("çok", "baya", "en" etc.) which have a boosting effect when met before an adjective. We handle punctuations like '!' after sentimental terms as boosters but giving less strength.

Instead of with/without words in English we have absence/presence suffixes (**+sız/+siz (without), +lı/+li (with)**) in Turkish which are added to nouns and change their POS tag to adjective. It is a kind of negation and changes the polarity of the following word. If any absence/presence suffixes met we do not eliminate these suffixes (“umut-suz"). As we mentioned before we also add these sentimental adjectives with absence/presence suffixes to the lexicon. So in sentiment score calculating process we compare these words with Lexicon.

The ML approach treats the sentiment analysis as a supervised classification problem. Supervised classification requires a sufficently large labeled dataset for proper training but Lexicon based sentiment analysis does not. Determining of feature set is another key process for ML classification. In order to create the feature vector, we use unigrams and bigrams by using inverse-document-frequency (TF-IDF) feature ranking and selection method. We conduct our experiments using SVM, NB and Decision Trees (J48) classification algorithms. 10 fold cross validation technique is utilized to train and test our supervised classifiers.

We use accuracy measure, the number of instances that predicted correctly, to evaluate performance of our systems. We activate and deactivate modules to show the contribution of each module to performance of sentiment analyzers.

According to results, each module has a contribution to the performance of Lexicon based sentiment analysis method but the most effective ones are Negation handling and MWEs handling for Twitter dataset and deasciification and negation handling for Movie dataset. The performance of Lexicon based sentiment analysis Method is 75.2% for Twitter dataset and 79.5% for Movie dataset. Results show that MWEs extraction and handling absence/presence suffixes bring reasonable improvement to performance of Lexicon based method. Since Movie reviews are too long and have too many sentimental words, MWEs extraction option does not bring enough improvement.

As most researchers [1] [8] reported, our results also show that SVM hashighest accuracy than other algorithms for ML approach. The best performance of ML Based Sentiment Analysis Method is 85.0% (SVM) for Twitter dataset and 89.5% (SVM) for Movie dataset. Using unigrams and bigrams together gives the best performance for almost all classifiers on both datasets. Results indicate that bigrams can handle most of consecutive cases such as negation, boosting and MWEs.

As surface forms of words include enough linguistic information such as negation and absence/presence suffixes, the usage of surface forms that combined with unigrams increases the performance of ML based method slightly for Movie dataset. But it decreases the performance for Twitter dataset since Twitter dataset is too noisy and feature selection threshold leaves most of bigrams below the feature selection threshold (min. 20 occurrence in Movie dataset and min. 5 occurrence in Twitter dataset). It decreases the performance when combined with unigrams+bigrams for Movie dataset also.

In comparison of these two methods, ML based method performs better than Lexicon based method on both short (Twitter dataset) and long informal texts (Movie dataset). The results show that accuracy of Movie dataset is better than accuracy of Twitter dataset in both Lexicon based and ML based sentiment analysis methods. Although Lexicon based sentiment analysis is unsupervised, it works well when text does not include sarcasm or irony.

1. GİRİŞ

1.1 Motivasyon

Sosyal bir varlık olarak biz insanlar, çevremizle etkileşim içindeyizdir. Bu sosyalliğin bir sonucu olarak başkalarının neler düşündüğünü merak eder ve öğrenmek isteriz. Merak ettiğimiz düşünceler bazen genel bazen de daha özel bir konu (kişi, nesne veya ideoloji) hakkında olabilir. Günlük hayatta birçok konuda yakın çevremize danışma ihtiyacı hissederiz. Bu danışma ihtiyacı, hakkında yeteri kadar fikir sahibi olmadığımız konularda olabileceği gibi içinde bulunduğumuz sosyal çevredeki görüşlerin bizim düşüncemize ne oranda örtüştüğünü öğrenmek için, bildiğimiz bir konu hakkında da olabilir. Bir ürün almak veya bir sinema filmi izlemek istediğimizde öncelikle bu ürün veya film hakkında fikri olan kişilere danışma ihtiyacı hissederiz. Öncelikle bize en yakın insanlara, onlar yetmezse o ürün hakkında daha fazla bilgi ve tecrübesi olan kişilere veya o alanda hizmet veren kurumlara başvurmak durumunda kalırız. Böylece yanlış bir ürünü almamış ve izlemeye değmeyecek bir film için zaman harcamamış oluruz. Bu çoğu kez bize parayı ve zamanı daha verimli kullanabilmemizi sağlar.

İnsanların belli konular ve ürünler hakkında neler düşündüğü, nelerden hoşlanıp nelere ihtiyaç duyduğu ticaret, üretim ve hizmet sektörleri için de ilgi çekicidir. Firmalar kendileri, ürünleri veya hizmetleri ile ilgili genelde tüm halkın, özelde ise hedef kitlelerinin neler düşündüğünü bilmek ve bunlara göre pozisyon almak isterler. Halk arasında son dönemlerde nelerin moda olduğu, nelerin sempati ile karşılandığı bilgisi firmaların ürünlerinde ve reklamlarında bu konuları uygun bir şekilde kullanabilmelerine yardımcı olabilir. Örneğin, sinema sektöründe çekilecek yeni bir filmin hasılat ve reyting açısından başarılı olması için benzer türde filmler ve oynatılacak oyuncular hakkında sinema izleyicilerinin düşüncelerini göz önünde bulundurulabilir. Bunun için izlenme reytingleri veya kullanıcı yorumları kullanılabilir. Siyasi partiler ülkenin genel sorunları, kendi partilerinin veya diğer partilerin manevralarına ilişkin halkın düşünce ve tepkisini öğrenip buna göre yeni

siyasi politikalar üretebilir ya da var olan politikalarında değişikliğe gidebilirler. Şirketler veya siyasi partiler tüm bunları yapabilmek için genellikle kayda değer paralar karşılığında anketler yaptırırlar.

Tüm bu ihtiyaçların giderilme şekli internetin gelişmesi ile beraber geleneksel yolların dışına çıkmıştır. İnternetin yaygınlaşmasıyla beraber insanlar belli konular hakkındaki görüşlerini forumlar, bloglar ve sosyal medya platformlarında paylaşmaya başlamışlardır. Bu alanlarda hızla biriken veri ve bu verilere kolay ulaşım imkânı, araştırma, sosyal analiz ve anketler için yeni bir adres oluşturmuş durumdadır. Artık bir kitap okumak istediğimizde o kitabı okumuş birilerine sormamıza gerek kalmadan internet üzerinden o kitapla ilgili yorumlara, eleştirilere ve o kitabın aldığı reytinge (satış rakamları) bakarak okuyup okumamaya karar verebiliriz.

Bu yöntemin büyük miktarlarda veriye uygulanmasıyla toplumun veya hedef kitlenin ihtiyaçlarını, genel yönelimlerini ve belli bir kişi, olay ve bunların özellikler hakkında duygusal analizini elde etmek mümkündür. Bu analizin otomatik olarak bilgisayarlar tarafından ve belli teknikler kullanarak yapılması piyasada kullanılabilecek araçların oluşmasına olanak sunar. Bu araçlar, firmaların kendileri, ürünleri ya da hizmetleri ile ilgili genelde tüm halkın özelde ise hedef kitlelerin neler düşündüğünü öğrenmesine ve bu istatistikleri kullanarak yeni hamleler yapabilmelerine olanak sağlar.

Bu veri analiz metotlarından biri olan Duygu analizi (DA), herhangi bir yazı içindeki kelime ve kelime öbeklerini kullanarak o yazının barındırdığı duyguyu çıkarmaya çalışır.

Son yıllarda Türkçe için yapılan birkaç DA çalışması olmasının yanında bugüne kadar yapılan çalışmaların çoğu İngilizce için yapılmıştır. Bu çalışmadaki amaç farklı Türkçe veri kümeleri üzerinde daha önce İngilizce ve Türkçe için yapılan çalışmalardaki makine öğrenmesi ve sözlük tabanlı DA çalışmalarına benzer iki sistemin yaratılıp başarımlarının hesaplanması ve karşılaştırılmasıdır. Bu amaçla iki farklı karakteristikte veri kümesi yaratılmış ve değerlendirme için kullanılmıştır.

Ayrıca performansı arttırabilecek, daha fazla bilgi çıkarımı sağlayacak yeni özelliklerin ve metotların incelenmesi de bu tezin amaçları arasına girmektedir. Bunlar kısaca Türkçe'nin karakteristik özelliklerine uygun ön işlemler, yeni öznitelik çıkarımları ve bilgi çıkarımı yöntemlerinin kullanılması olarak belirtilebilir.

Bileşik kelimelerin çıkarılması ve varlık/yokluk eklerinin ele alınması gibi yeni modüllerin DA'ne katkısının araştırılması bu tezin başka bir amacıdır.

1.2 Duygu (Sentiment) Analizi Nedir?

İnsanlar konuşurken ya da yazarken genel olarak iki sınıfta kategorize edilebilen ifadeler kullanırlar: gerçekler ve görüş bildiren ifadeler. Gerçekler kişiler, olaylar ve bunların özellikleri hakkında nesnel ifadeler bildirirken, görüşler genellikle özneldirler ve belli konular hakkında insanların duygularını, görüşlerini veya değerlendirmelerini içerirler. Bu değerlendirmeler yazarın o anki psikolojik durumu, okurda bırakmak istediği etki ve ilgili konu veya şahıs hakkındaki tutumuna bağlı olarak değişebilmektedir. Görüş kavramı çok kapsamlı olmakla beraber belli bir kişi, olay ve bunların özellikleri hakkında olumlu, olumsuz veya nötr ifade içeren olmak üzere üç sınıfa ayrılabilmektedir [2].

Psikoloji bilim dalı DA'ni çokça işlemiş ve kişinin duygusal durumunun, kullandığı kelimeler ve bu kelimeleri kullanma şekilleriyle çok yakından ilgili olduğu tespit edilmiştir [16]. Bu amaçla duygu barındıran kelimeler birçok çalışmada duygu yönelimlerine göre sınıflandırılmış ve taşıdıkları duygu yoğunluğuna göre puanlandırılmışlardır. Daha sonra uygulanan seanslarda hastaların bu kelimeleri kullanma şekilleri ve kullanma sayıları tedavi sürecinde kullanılmıştır [16]. Aynı şekilde sosyal araştırmalarda yapılan anketler ve incelemelere bakarsak DA'nin sosyoloji bilim dalını da yakından ilgilendirdiğini görürüz. Örnek vermek gerekirse toplumsal tepkilerin ve ayaklanmaların yoğun olduğu dönemlerde toplumsal olarak genelde ya da özelde belli konularda olumsuz bir dil kullanırız. Bu tepkiyi doğuran kişi veya kurum hakkında yapılacak kapsamlı bir anket çalışması bize toplumun o konu hakkında duygusal durumunu gösterebilmektedir.

Günümüzde DA hem bireyi hem de toplumu incelemek için önemli bir veri kaynağı sunar. Büyük kitlelerin duygusal tepki ve yönelimlerini inceleyebilmek, o kitleleri hedef gözeten kurum ve kuruluşlar için müthiş bir kaynak oluşturmaktadır. Toplumun belli bir kişi, ürün veya konu hakkında neler düşündüğü, genel olarak neleri sevip sevmediği, nelere ihtiyaç duyduğu gibi veriler üretim ve hizmet sektörleri için önemlidir.

1.3 Tezin Organizasyonu

Bu bölümde DA'nin tanımı, önemi, motivasyonu, ilgili çalışmalar ve bu çalışmanın özerinde durulmuştur. Bölüm 2'de, DA ile ilgili daha önce yapılmış çalışmalar; yaklaşımları ve sonuçlarıyla birlikte verilmektedir. Bölüm 3'te, bu çalışmada kullanılan yöntemlerin bilimsel altyapısı ve dayanakları irdelenmektedir. Bölüm 4'te, bu çalışmada kullanılan veri kümeleri ve metotlar detaylı bir şekilde sunulmaktadır. Bölüm 5'te, deneysel çalışmalar ve sonuçları verilip tartışılmaktadır.

2. LİTERATÜR ARAŞTIRMASI

Duygu analizi bir sınıflandırma problemidir. DA ile ilgili Makine Öğrenmesi (MÖ) ve sözlük tabanlı yöntemlerle birçok akademik çalışma yapılmıştır.

Forum, blog ve sosyal medyanın katkısıyla internet ortamında biriken bilgi miktarı hızla artmaktadır. Bu büyük veri içinde çok miktarda duygu barındıran bilgi de bulunmaktadır. Bu bilgiye ulaşmak, işlemek, öznelliği ortaya çıkarmak ve duygu barındıran ifadeleri sınıflandırmak, DA'nin temel amaçlarını oluşturmaktadır. İğneleme ve ironi hem psikoloji [3] hem de DDİ [4] [17] alanında büyük öneme sahiptir ve fazlasıyla ilgi çekici bir konumdadır. Doğal bir metindeki iğneleme ve ironinin anlaşılması insanlar için bile zor bir durumdur [4]. İğneleme ve ironinin yakalanmasındaki başarı artışı, DA'nin de başarımını önemli ölçüde artıracağı görülmektedir.

DA problemi için akademik birçok çalışma yapılmıştır. Bunlardan çoğu öznellik çıkarımı ve duygu durumu sınıflandırmaya yoğunlaşmaktadır [2] [7]. Sınıflandırma için çoğunlukla sözlük ve MÖ tabanlı yaklaşımlar kullanılmaktadır. Özellikle son yıllarda DDİ ve Görüntü İşleme alanlarında yüksek başarımlı sonuçlar veren derin sinir ağları tabanlı Derin Öğrenme(DÖ) yöntemi de DA için kullanılan önemli yöntemlerden birisidir. Bu yöntem İngilizce için çokça kullanılan ve literatürde en yüksek başarımların elde edildiği çalışma alanı olarak karşımıza çıkmaktadır [18] [19].

Pang ve diğ. [7], çalışmalarında DA problemini konu bağımsız metin sınıflandırması olarak ele almışlardır ve performanslarına göre karşılaştırmak üzere değişik MÖ teknikleri uygulamışlardır. IMDB adlı sinema değerlendirme platformundan aldıkları film yorumlarını olumlu-olumsuz sınıflandırmaya tabi tutmuşlardır. Bu çalışmalar sonucunda KDM ile %82 en yüksek başarılarını elde etmiş ve DA sınıflandırmanın normal konu tabanlı doküman sınıflandırmasına göre daha zor bir konu olduğu sonucuna varmışlardır.

Jiang ve diğ. [20], tweetler üzerinde hedef-bağımlı (target-dependent) bir DA sınıflandırma uygulamışlardır. Tweetler üzerinde hedef-bağımsız bir DA'nin, ürün ve film yorumlarında olduğu gibi doğru bir yaklaşım olmadığını, tweetler genellikle ilgili hedefin yanında başka hedefler de barındırdığdan, hedef-bağımlı bir yaklaşımın daha doğru olacağını belirtmişlerdir. Ayrıca tweetlerin çoğu kez kısa olmasından (140 karakter) dolayı ilgili hedef hakkındaki duyguyu yakalamak çok zorlaşmaktadır. Bunun için Jiang ve diğ., bağlamın (ilgili tweetlerin) da dikkate alınması gerektiğini belirtmişlerdir. Jiang ve diğ., sınıflandırma için linear kernel ile SVM-Light sınıflandırıcısını kullanmışlardır. Bir tweetteki farklı hedefleri ayırd edebilmek için POS etiketleri, kelime kökü, biçimbirimsel çözümleme gibi temel bazı DDİ teknikleri kullanmışlardır. Jiang ve diğ., hedef-bağımlı ve bağlam duyarlı özniteliklerin kullanılmasıyla elde ettikleri %85.6 başarımlı DA metot ile, tweetler ve benzeri veri kümelerinde, bu özelliklerin ne kadar önemli olduklarını göstermişlerdir.

Turney [21], anlamsal yönelimlerine göre yorumları tavsiye edilebilir veya tavsiye edilemez olarak sınıflandırmak için basit bir denetimsiz öğrenme algoritması uygulamıştır. "Excellent (harika)" ve "poor (kötü)" gibi kelimeler ile sınıflandırılmak istenen yorumlardaki kelimelerin ortak bilgilerini kullanarak o yorumların duygusal yönelimlerini belirlemeye çalışmıştır. Bu çalışmada da, MÖ sınıflandırma çalışmalarının büyük çoğunluğunda olduğu gibi, farklı özelliklerin birleşiminden yeni özellikler yaratma kabiliyetine sahip olan KDM algoritması en iyi sonucu vermiştir.

Bo Pang ve Lillian Lee [22], katmanlı sınıflandırıcı mantığıyla önce veriyi öznel-nesnel olarak sınıflandırmış daha sonra öznel bulunanları olumlu-olumsuz olarak sınıflandırmışlardır. 10000 yorum (5000 olumlu, 5000 olumsuz) kullanılarak yapılan çalışmalarında bir önceki çalışmalarına göre iki sınıflı sınıflandırmada %4 lük bir artışla %86 başarı sağlamışlardır.

Nguyen ve diğ. [23], yaptıkları çalışmada twitter verisini analiz ederek önceki tweetlerdeki algıyı kullanıp zaman içerisindeki algı değişimine bağlı olarak gelecek tweetlerdeki algıyı tahmin etmeye çalışmışlardır. Twitterin dinamik yapısını en iyi belirleyen öznitelikler seçilerek; KDM, lojistik regresyon ve karar ağaçlarının (KA) kullanıldığı çalışmada en yüksek başarımı veren KDM, %85 bandında bir başarı ile öne çıkmaktadır.

Socher ve diğ. [18], anlamsal kelime uzaylarını kullanılan yöntemlerin uzun ifadelerde başarılı olamayacaklarını, bunun için daha güçlü denetimli öğrenme sunan yöntemlerin gerektiğini belitmişlerdir. Bu yöntemlerin en umut verici olanının da DÖ (Derin Öğrenme) olduğunu yaptıkları çalışmayla göstermeye çalışmışlardır. Çalışmalarında geniş, duygusal olarak etiketli kelimeler içeren cümlelerden oluşan ağaç yapılı bir derlem oluşturmuşlardır. Bu derlemi kullanarak yinelemeli DÖ ile olumlu/olumsuz DA senaryosunda %85.4 başarım elde etmişlerdir.

DA uygulamaları ve yaklaşımlarının büyük çoğunluğu İngilizce için yapılmasına rağmen son dönemlerde diğer diller için de DA yaklaşımları ve uygulamalırının geliştirilmesi popüler olmuştur. Özellikle DA konusu sosyal medyanın müthiş ilerleyişiyle birlikte daha büyük önem kazanmış ve her dil için talep edilir duruma gelmiştir. Özellikle DA'nin ticari olarak piyasa araştırmalarında kullanılabilir oluşu DA'nin bütün diller için gerekliliğini ortaya koymaktadır. Türkçe için çok az sayıda DA ile ilgili çalışma olduğundan, bu konu araştırmaya açık ve dikkat çeken bir konudur.

Türkçe için yapılan çalışmalardan biri Eroğul'un yüksek lisans tez çalışmasıdır [8]. Eroğul tezinde DA problemini bir sınıflandırma problemi olarak ele alıp, değişik MÖ yaklaşımları uygulayarak performanslarına göre karşılaştırmaktadır. Çalışmasında film yorumlarını değerlendiren Eroğlu olumlu-olumsuz sınıflandırma probleminde %85 başarı elde etmiştir.

Vural ve diğ., Türkçe film yorumları için sözlük tabanlı bir DA çalışması [6] yapmışlardır. Sentistrengths kütüphanesini Türkçeye çevirerek DA problemine çözüm bulmaya çalışmışlardır. Onlar da Eroğlu'nun kendi çalışmasında kullandığı ve "beyazperde.com" adresinden topladığı veri kümesini kullanmışlardır. Çalışmalarında olumlu-olumsuz sınıflandırma senaryosunu işlemiş ve %76 başarı elde etmişlerdir.

Meriç ve Diri'nin Twitter verisi üzerinde yaptıkları DA [24] çalışması da diğer önemli çalışmalardan biridir. Çalışmalarında MÖ yöntemini denetimli sınıflandırıcılar ile uygulamışlardır. Alan (domain) bağımlı ve alan bağımsız veri kümelerine uyguladıkları sözcük tabanlı ve 2 ve 3 karakter n-gramlı yaklaşımlarla, bu yaklaşımların ilgili veri kümesi türlerinde denetimli sınıflandırıcılar ile sağladıkları başarımların karşılaştırılmasını hedeflemişlerdir. Çalışmaları sonucunda sözcük

tabanlı denetimli sınıflandırmanın alan bağımsız veri kümelerinde, karakter n-gram tabanlı denetimli sınıflandırmanın ise alan bağımlı veri kümelerinde daha başarılı olduğunu görmüşlerdir.

Şimşek ve Özdemir, çalışmalarındada [25] borsadaki değişim ile Twitter kullanıcılarının ekonomi ile ilgili attıkları tweetler arasında bir ilişki olup olmadığını araştırmışlardır. Duygu sözlüğünden sekiz farklı duyguya (öfke, hüzün, aşk, korku, iğrenme, utanç, eğlence, sürpriz) ait 113 özellik seçilerek, bu özellikler ışığında tweetler mutlu-mutsuz olarak sınıflandırılmıştır. Yapılan çalışma sonucunda borsadaki değişimlerin tweetlerin mutlu-mutsuz olma durumlarıyla %45 ilişkili olduğu saptanmıştır.

3. BİLİMSEL ARKA PLAN

Duygu analizi (DA) konusu, Makine Öğrenmesi (MÖ), Doğal Dil İşleme (DDİ) ve Bilgi Çıkarımı (BÇ) konularıyla yakından ilgilidir. MÖ, belli özniteliklere göre sınıflandırma noktasında DA'ni bir sınıflandır problem olarak ele alır. Karar Destek Makineleri (KDM) ve Naive Bayes (NB), MÖ alanından sınıflandırma için kullanılan araçlardır. DDİ, kelimelerin biçimbirimsel analizinin yapılması, sözcük türlerinin belirlenmesi ve belirsizlik giderimi konularında kullanılmaktadır. Bilgi çıkarımı ise her kelimenin metin ve veri kümesi içerisindeki bulunma istatistiklerine göre kelime-sınıf ilişkisi yaratma ve noktasında kullanılmaktadır. Kelimelerin istatistiksel özellikleri olarak Term Frequency-Inverse Document Frequency (TF-IDF) teknikleri kullanılmaktadır.

3.1 Makine Öğrenmesi

Makine öğrenmesi bilgisayarlara, programlama yapılmadan, öğrenme yeteneği sağlayan bir yapay zekâ tekniğidir. MÖ, kendilerini geliştirmek için eğitebilen, yeni veriler ile kendilerini değiştirebilen bilgisayar programlarının geliştirilmesi üzerinde durur. Bilgisayarlara karmaşık örüntüleri algılatma ve veriye dayalı akılcı kararlar verebilme becerisi kazandırmak, MÖ araştırmalarının odaklandığı konudur. MÖ, istatistik, olasılık kuramı, veri madenciliği, örüntü tanıma gibi alanlarla yakından ilintilidir.

MÖ süreci, veri madenciliği sürecine benzer. Her iki sistem de desenleri aramak için veri üzerinde tarama yapar. Buna karşın, veri madenciliği veriyi insanların karşılaştırıp bilgi çıkarabilmeleri için elde ederken, MÖ elde ettiği bilgiyi programın kendi öğrenme becerisini geliştirmesi için kullanır.

MÖ'nin başlıca uygulama alanları, makine algılaması, bilgisayarlı görme, doğal dil işleme, sözdizimsel örüntü tanıma, arama motorları, tıbbi tanı, biyoinformatik, kredi kartı dolandırıcılığı denetimi, borsa çözümlemesi, DNA dizilerinin sınıflandırılması, konuşma ve elyazısı tanıma, bilgisayarlı görmede nesne tanıma, oyun oynama gibi

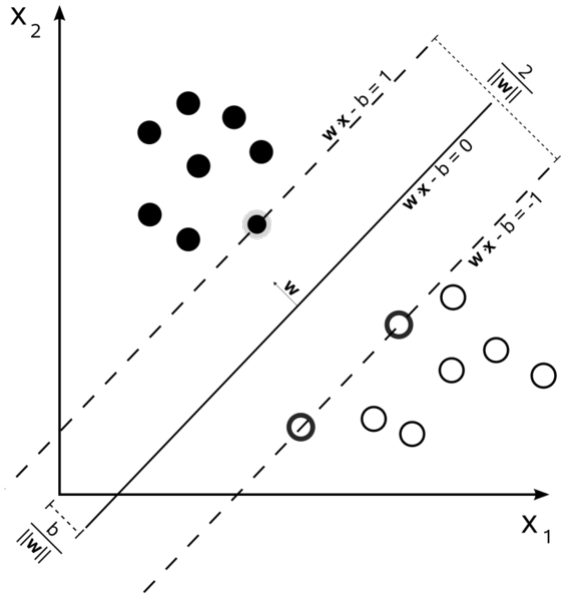
sıralanabilir. MÖ'nin birçok alanda başarılı sonuçlar vermesi, DDİ için kullanılmasını da popülerleştirmiştir.

MÖ denince, akla ilk olarak sınıflandırma ve sınıflandırıcılar gelir. Genel olarak bir sınıflandırma problemi, MÖ'nde denetimli veya denetimsiz öğrenme algoritmalarıdır. Denetimli sınıflandırma yapılırken öncelikle hangi sınıfa ait oldukları belli, önceden etiketli yeterince büyük bir eğitim kümesinin olması gerekir. Denetimli sınıflandırma algoritması (Naive Bayes, KDM, Karar Ağaçları (KA) vb.) bu eğitim kümesindeki örüntüleri öğrenerek bir model üretir. Artık bu model istenilen etiketsiz örnekleri, eğitim kümesinden öğrendiği örüntülere göre, sınıflandırabilir. E-posta kutusuna gelen e-postaların spam olarak ayrıştırılması işlemi buna örnek verilebilir. Bu örnekte spam e-posta ve spam olmayan e-posta ayrıştırılacak iki sınıfı temsil eder. Elimizdeki spam ve spam olmayan e-postalardan yaralanarak, bu iki sınıfın özelliklerine göre gerekli örüntüleri öğrenip, bu bilgilere göre gelecekte elimize ulaşacak e-postaların spam olup olmadığına karar verecek bir algoritma denetimli makina öğrenmesine örnektir.

Bu çalışmada MÖ tekniklerinden karar destek makineleri sınıflandırıcı ve naive bayes sınıflandırıcı kullanılmıştır. Karar destek makineleri daha önceki birçok çalışmada en iyi başarımı sağlayan teknik olmakla beraber naive bayes sınıflandırıcı bazı alanlarda ve ya veri kümelerinde gayet başarılı sonuçlar verebilmektedir.

3.1.1 Karar Destek Makineleri

Karar Destek Makineleri (KDM) ayırıcı bir hiperdüzlem ile tanımlanabilecek ayrıştırıcı bir sınıflandırıcıdır. Girdi olarak her biri farklı iki kategoriye etiketli veri verildiğinde (denetimli öğrenme), KDM eğitim algoritması çıktı olarak, verilecek yeni etiketsiz örnekleri sınıflandırabilecek bir hiperdüzlem (model) üretir. MÖ'ne girdi, ilgili veriyi reel sayılarla ifade eden bir öznitelik vektörü olarak verilir. Veriyi reel sayılardan oluşan vektörler olarak ifade etmek, veri kümesine ve veri türüne göre zor bir işlem olabilir. Metin işleme alanında bir metni öznitelik vektörüne dönüştürmek için bag-of-words metodu kullanılabilir. Bag-of-words metodunda her kelime öznitelik vektörünün bir elemanı olarak yer alır. Öznitelik vektöründe kelimeleri ifade etmek için reel bir sayı belirlenirken; o metin içerisindeki frekansı (TF), o metinde bulunup bulunmama durumu (binary) ya da tüm eğitim setindeki frekansına göre (DF) vb. değerler hesaplanıp kullanılabilmektedir.



Şekil 3.1: Karar Destek Makineleri (KDM) çalışma prensibi ve maksimum margin.

Bir KDM modeli, örneklerin sahip oldukları öznitelik değerlerine göre uzayda noktalar olarak ifade edilmiş durumlarıdır. Bu model, örnekleri kategorilerine göre ayırır. Bunu yaparken kategoriler arasındaki mesafeyi (margin) olabildiğince büyük yaparak toplam hatayı minimize etmeye çalışmaktadır [26]. Gelecek olan yeni örnekler, model uzayındaki kategorilere göre ayrılmış bölgelere düşüp düşmemelerine göre kategorize edilmektedir.

Gerçek problemler genellikle çok boyutlu uzayda yer alır ve ayrılması gereken gruplar doğrusal olarak ayrılabilir olmayabilirler. KDM doğrusal sınıflandırmanın yanı sıra farklı kernel fonksiyonları ile doğrusal olmayan sınıflandırmaları da girdileri daha yüksek boyutlu bir öznitelik uzayına yükselterek başarılı bir şekilde icra edebilmektedir [27]. Kernel oyunu birkaç özniteliğin birleşiminden yeni öznitelikler yaratma şeklinde gerçekleştirilmektedir. Kernel fonksiyonları, KDM'nin yüksek performans göstermesinin en önemli etkenlerindendir.

Denetimli MÖ için, yeterli sayıda, ait oldukları sınıfa göre etiketli örnek içeren bir eğitim kümesine ihtiyaç vardır. Bir eğitim kümesi (Eşitlik 3.3) ile ifade edilebilir. Sınıfların birbirlerinden ayrılmasını sağlayan ve belirli yöntemlerle seçilen öznitelikler (Eşitlik 3.1), öznitelik vektörünü (Eşitlik 3.2) oluşturur. Eğitim kümesi ve öznitelik vektörü sınıflandırıcıya verilir. KDM sınıflandırıcı bu öznitelikleri kullanarak, örnekleri sınıflara göre birbirlerinden bir hiperdüzlem (Eşitlik 3.4) ile ayırmaya çalışır (Şekil 3.1). Bunu yaparken, sınıflar arasındaki mesafeyi maksimum yapacak w ve

b değerlerini bulmaya çalışır [27]. Burada w hiperdüzleme dik normal doğrusu iken b hiperdüzlemin orjine olan uzaklığıdır ve örneklem uzayının biased/unbiased durumunun ölçüsüdür [27].

$$x(i), i = 1,2,...,L \tag{3.1}$$

$$x = \left[x(1), \quad x(2) \quad \cdots \quad ,x(L)\right]^T \in R^L \tag{3.2}$$

$$D = \{(x_i, y_i) \,|x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^n \tag{3.3}$$

$$\mathbf{w.x} - b = 0 \tag{3.4}$$

3.1.2 Naive Bayes sınıflandırıcı

Naive Bayes bağımsızlık önermesini kullanan basit bir istatistiksel sınıflandırıcıdır. Bu önerme "sınıf koşullu bağımsızlık" olarak adlandırılır ve sınıflandırmada kullanılacak her bir öznitelik ya da parametrenin istatistiksel açıdan bağımsız olması gerekliliğini ifade eder. Daha açık bir ifadeyle verilen bir sınıf etiketine herhangi bir özniteliğin etkisi diğer özniteliklerin var olup olmamasına bağlı değildir. Bir diğer deyişle Naive bayes sınıflandırıcısına bayes teoreminin bağımsızlık önermesiyle basitleştirilmiş hali diyebileceğimiz gibi "bağımsız öznitelik modeli" diye de tanımlayabiliriz. Basit bir örnek vermek gerekirse bir arabanın spor araba olabilmesi için şu özelikler önemlidir: "motor gücü", "hız üst sınırı", "tork" ve "fren" değerleri. Bir naive bayes sınıflandırıcısı bu özelliklerin her birinin, bir arabanın spor araba olup olmamasına olan katkısını, diğer özelliklerin olup olmamasını dikkate almaksızın ayrı ayrı ve birbirinden bağımsız olarak ele alır. Naive bayes sınıflandırıcısının avantajlarından biri diğer sınıflandırıcılara göre çok az miktarda eğitim kümesi ile gerekli parametreleri (değişkenlerin ortalama ve varyansı) tahmin edebilmesidir. Bunun nedeni, özniteliklerin bağımsızlığı sayesinde, tüm kovaryans matrisinin yerine sadece ilgili sınıfa ait değişkenlerin kovaryansı hesaplanıyor olmasıdır. Naive bayes algoritması her özniteliğin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır.

Eğitim kümesi ile eğitilen naive bayes sınıflandırıcısı, kullanılan özniteliklerin her birinin sınıflarla olan ilişkisini olasılık oranı olarak hesaplar ve o değerleri içeren modeli çıktı olarak verir. Daha sonra naive bayes sınıflandırıcısı, test örneklerini, bu modeldeki öznitelik-sınıf olasılıklarını kullanarak, özniteliklerin bağımsızlığı varsayımıyla sınıflandırır. Sınıflandırma işlemi şu şekilde gerçekleştirilir: 3.5, 3.6 ve 3.7 eşitliklerindeki $P(S_i)$ ve $P(S_j)$ sırasıyla sınıflandırılma yapılacak i ve j sınıflarının öncel olasılıkları, $P(S_i|x)$ ve $P(S_j|x)$, sırasıyla i ve j sınıflarının ardıl olasılıkları, P(x) x'in olasılık yoğunluk fonksiyonu ve $P(x|S_i)$ x'in i sınıfına bağlı koşullu olasılık yoğunluk fonksiyonu olsun. Bayes karar teoremine (3.6) göre x örneği sınıf i'ye aittir. eğe Bayes karar teoremine özniteliklerin istatistiksel olarak bağımsızlığı varsayımı eklenirse bir Naive bayes sınıflandırıcısı (Eşitlik 3.7) elde edilir. Bu durumda x örneğini tanımlayan bütün özniteliklerin sınıflara göre durumlarının katkıları bağımsız olarak işleme dahil edilir.

$$P(S_i|x) \times p(x) = p(x|S_i) \times P(S_i) \tag{3.5}$$

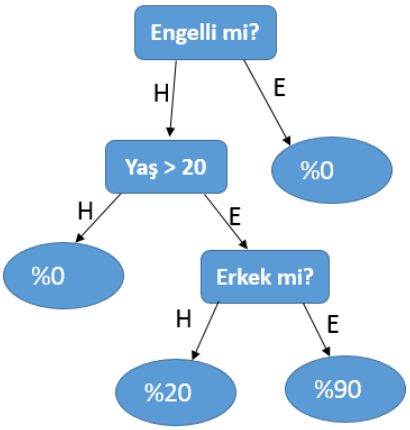
$$P(x|S_i)\,P(S_i) > p\left(x|S_j\right)P\left(S_j\right), \forall j \neq, i \tag{3.6}$$

$$P(S_i)\prod_{k=1}^{L} P(x_k|S_i) > p\left(S_j\right)\prod_{k=1}^{L} P\left(x_k|S_j\right) \tag{3.7}$$

Naive bayes sınıflandırıcının kullanım alanı bağımsızlık önermesinden dolayı her ne kadar kısıtlı gözüksede yüksek boyutlu uzayda ve yeterli sayıda veriyle öznitelik kümesi bileşenlerinin istatistiksel bağımsız olması koşulu esnetilerek başarılı sonuçlar elde edilebilinir[13].

3.1.3 Karar Ağaçları

Karar ağaçları (KA) sınıflandırma yöntemi, sürekli ve kesikli değerlerle çalışabilen açgözlü öğrenme temelli bir sınıflandırıcıdır. KA öğrenmesinde, bir ağaç yapısı oluşturularak ağacın yaprakları seviyesinde sınıf etiketleri ve bu yapraklara giden kollar ile de özellikler üzerindeki işlemeler ifade edilmektedir. Bu ağacın bütün iç düğümleri (interior nodes) birer girdiyi ifade eder. Karar ağaçları, denetimli olarak çalışan sınıflandırıcılardır ve son durumlar (yapraklar) hariç bütün durumlar (düğümler) belli bir özniteliğe göre bir kuralı ifade eder. Bu kural o karar aşamasından sonra seçilecek dalı belirler. Son durumlar sınıflandırmanın sonuçlarını (sınıf



Şekil 3.2: Karar ağaçları (KA) çalışma prensibi.

etiketlerini) taşır. Bu sınıflandırma sonucuna bütün karar düğümlerindeki kurallar ile ulaşılmaktadır.

Şekil 3.2'teki dikdörtgenler kural düğümlerini göstermektedir. Kural düğümleri, örneklerin öznitelik değerlerine göre oluşturulur. Bu karar ağacındaki öznitelikler yaş, cinsiyet ve fiziksel engelli durumlarıdır. Yuvarlak düğümler, karar ağacının yapraklarıdır ve sınıflandırıcı sonucunda elde edilen sınıf etiketlerini belirtir. Bu KA sınıflandırıcısında bir kişinin orduya alınıp alınamayacağı durumu belirlenmektedir. Burada varsayım olarak kadınların askere alınma olasılığının düşük olduğu kabul edilmiştir.

Düğümler arasındaki bağlantılar, özniteliklere göre izlenecek yolu göstermektedir. Karar ağaçlarında öznitelik seçimi önemlidir ve bu bilgi kazancı oranı (information gain ratio) değeriyle belirlenmektedir. Bilgi kazancı, hangi özelliğin sınıflandırma için en iyi ve avantajlı olduğu bilgisini verir. Bilgi kazancı oranı (BKO, Eşitlik 3.8), karar ağaçları öğrenmesinde bilgi kazancının (BK) asıl bilgiyi veren bir oranıdır. Bu değer, bir özellik seçerken, çok-değerli (multi-valued) bir özelliğe olan eğilimi (biased), dallanma sayısını ve boyutunu gözönüne alarak en aza indirmeye çalışır. Bilgi kazancı ise, veri setindeki örneklerin sınıflara göre dağılımının ne kadar düzgün olduğunu gösteren Entropi (Eşitlik 3.10) adında bir değer kullanır. Entropi, iki sınıflı bir veride, örnekler her sınıfa eşit miktarda dağılmışsa, yani her sınıfta eşit sayıda örnek varsa, minimum değer (0) alır. Eşitlik 3.11'deki $\text{Entropi}_A(D)$ değeri A özniteliğinin değerine

göre, örneklerin dağılımının entropisini ifade eder. Bir diğer deyişle belli bir noktadan doğan yeni durumların ortalama entropisidir.

Karar ağaçlarında az kuralla sonuca gitme amacı vardır (açgözlü yaklaşım). Bir özniteliğin değeri, belirleyiciliğine, yani ne kadar ayırdedici olduğuna bağlıdır. Diğer bir deyişle, karar düğümünde kullanılan öznitelik, bir örneği sınıflandırmada ne kadar kazançlı bir yol sağlıyorsa, o özniteliği kullanmak o kadar avantajlıdır ve iyi sonuç verir. Eşitlik 3.9'daki BK(A) ifadesi, bir A özniteliğinin bilgi kazancını ifade eder.

$$BKO(A) = \frac{BK(A)}{AB(A)} \tag{3.8}$$

$$BK(A) = Entropi(D) - Entropi_A(D) \tag{3.9}$$

$$Entropi(A) = -\sum_{i=0}^{n} p_i log_2(p_i) \tag{3.10}$$

$$Entropi_A(D) = -\sum_{k=1}^{y} \frac{|D_j|}{|D|} \, x \, Entropi(D_j) \tag{3.11}$$

$$AB_A(D) = -\sum_{k=1}^{y} \frac{|D_j|}{|D|} \, x \, log_2(\frac{|D_j|}{|D|}) \tag{3.12}$$

Bilgi kazancı, saflığı (Entropinin tersi) en yüksek olan, yani sınıflandırma sırasında en az bilgi gerektiren özniteliği seçme eğilimindedir. Sınıflandırma açısından istenmeyen bu durum, ayırma bilgisi (Eşitlik 3.12) sayesinde dengelenir.

Karar ağacı öğrenmesinde, ağacın öğrenilmesi sırasında, üzerinde eğitim yapılan küme, çeşitli özelliklere göre alt kümelere bölünür, bu işlem, özyineli olarak (recursive) tekrarlanır ve tekrarlama işleminin tahmin üzerinde bir etkisi kalmayana kadar sürer. Bu işleme özyinelemeli parçalama (recursive partitioning) ismi verilir.

Bir karar ağacının analizini yapabilmek için ağacın en son durumlarından en başa doğru bir değerlendirme yapılmalıdır. Beklenen değerler, incelenen karar aşamasındaki olasılık değerleri ile işlemin sonucunda ulaşılacak olan ödeme değerleri ile ağırlıklandırılarak toplanır.

3.2 Doğal Dil İşleme

Doğal Dil İşleme (Natural Language Processing) kısaltması olan DDİ (NLP) olarak bilinen, bilgisayar bilimi, bapay zekâ, dil bilimi alt kategorisi olan ve bilgisayarla insan (doğal) dillerinin etkileşimini inceleyen bu çalışma alanı, insan-bilgisayar etkileşimi çalışma alnı ile de yakından ilgilidir. DDİ, doğal dillerin kurallı yapısının çözümlenerek işlenebilmesi, anlaşılması veya yeniden üretilmesi amacını taşır ve otomatik çevri, soru-cevap makineleri, konuşma tanıma, konuşma üretme, metin özetleme, duygu analizi (DA) gibi birçok konudaki çalışmalarda kullanılmaktadır. DDİ, biçimbirimsel çözümleme, konuşma segmentasyonu, part-of-speech (POS) etiketleme, anlam belirsizliği giderme gibi birçok seviyede problemi çözmeye çalışır. Duygu analizi çalışmamızda kullandığımız biçimbirimsel analiz, biçimbirimsel belirsizlik giderme ve pos etiketleme metotları bu DDİ'nin kapsamına girmektedir.

3.2.1 Biçimbirimsel Çözümleme (Morphological Analysis)

Biçimbirimsel çözümleme, cümle içerisindeki her kelimenin kök ve eklerine ayrıştırılması ve görevlerinin belirlemesi sürecidir. Biçimbirim, dilde tek başına anlamı olmayan ancak kelime içerisine girdiği zaman anlam kazanan en küçük dilsel birimlerdir. Biçimbirimsel çözümlemede kelimelerin kök ve eklerinin çözümlenmesi ile beraber kelimelerin tipi (isim, fiil, sıfat, zarf, edat gibi) de belirlenir. Özellikle Türkçe ve Fince gibi sondan eklemeli dillerde bir kelimenin kökünden çok sayıda kelime türetilebildiğinden, biçimbirimsel çözümleme yapılması önemlidir. Biçimbirimsel Çözümlemede sözlük, imla kuralları, biçimbirimsel kuralları gibi girdiler ile isim soylu, fiil soylu kelimeler ve sayılar için tasarlanmış sonlu durum makineleri (SDM) kullanılarak sonuca ulaşılır.

Bu çalışmada Kemal Oflazer'in Türkçe biçimbirimsel analiz kütüphanesi [9] kullanılmıştır. Kullanılan biçimbirimsel analiz kütüphanesinin çalışma şekli Çizelge 3.1'te görülmektedir.

Çizelge 3.1: Biçimbirimsel çözümleyici çalışma şekli.

Anlamıyorum -> anla + Verb + Neg + Prog1 + A1sg					
Türkçe			**İngilizce**		
Anlamıyorum			I don't understand		
Kök	Fiil	anla+mak	Root	Verb	understand
Ekler	Fiil Kökü	Verb	Affixes	Verb Root	understand
	Olumsuzluk	Neg		Negation	Not
	Şimdiki zaman	Prog1		Simple Present Tense	Do/Does
	1. Tekil Şahıs	A1sg		1. Person Singular	I

3.2.2 Biçimbirimsel Belirsizlik Giderme

Biçimbirimsel belirsizlik giderme, biçimbirimsel çözümleyicinin cümle içerisindeki her kelime için verdiği birçok sonuçtan doğru olanı bulmak olarak ifade edilebilir. Türkçe, Fince ve Macarca gibi sondan eklemeli ve çekimli diller, karmaşık biçimbirim gibi özelliklerinden dolayı, DDİ için zor dillerdir. Karmaşık biçimbirimsel yapıdan doğan belirsizlik nedeniyle biçimbirimsel çözümleyici farklı kök ve biçimbirim (morfem) sıralamasına sahip birden çok çözümleme sonucu verebilir. Bu çalışmada Sak ve diğ. [10] 'nin Türkçe için biçimbirimsel belirsizlik gidericisi kullanılmıştır.

3.2.3 POS etiketleme

Part-of-Speech (POS), bir kelimenin dâhil olduğu dilbilimsel kategoriyi ifade eder. Part-of-speech etiketleme, cümle içerisindeki her kelimenin ayrıştırılıp, hangi dilbilimsel gruba dâhil olduğunu belirleme sürecidir. Dilbilimsel kategorileri kabaca şu şekilde sıralayabiliriz: "isim", "fiil", "sıfat", "edat", "zamir", "zarf", "bağlaç" ve "ünlem". POS etiketleme işlemi biçimbirimsel çözümleme içerisinde yapılan bir işlemdir. Dilbilimsel kategorilerin bilinmesi ve kullanılması, birçok DDİ probleminde olduğu gibi DA çalışmalarında da performansa ciddi bir şekilde katkı sunar [1] [2] [8]. Türkçe için mevcut biçimbirimsel çözümleyici ve dilbilimsel kategori etiketleme araçları [9] [28], birçok seviyede önemli analizler verebilmektedir.

Duygu analizi yapılırken, kelimelerin farklı dilbilimsel kategorilerde farklı anlamlar taşıması durumu önemlidir ve bunun yakalanıp kullanılması başarıma katkı sunmaktadır [1] [2] [8]. Örneğin; "ada (isim)" herhangi bir duygu barındırmazken

"ada+mak (fiil)" olumlu anlam içermektedir. Bu çalışmada, bu bilgiyi saklayıp kullanabilmek ve başarımdaki etkisini ölçebilmek için fiil tipin "fiil_kökü+eylem" şeklinde işaretlenmiştir.

3.3 Makine Öğrenmesi'nde DDİ

DDİ ve metin madenciliği gibi alanların kelimelerle, MÖ'nin ise reel sayılarla çalışıyor olması, bu alanlarda MÖ kullanılırken sorunların ortaya çıkmasına neden olmaktadır. Bunun üstesinden gelebilmek için MÖ için belirlenecek olan reel değerli öznitelikler oluşturulurken metindeki kelimeler, kelime sayıları, kelime türleri, n-gram'lar vb. özellikler kullanılmaktadır.

Burada önemli olan nokta özniteliklerin belirlenmesi sürecidir. Metinleri öznitelik vektörüne dönüştürme sürecinde, veri kaybının en aza indirgenmesi için seçilecek öznitelik setinin veri kümemizle ilgili olabildiğince fazla bilgi içermesi, ilgili çalışma alanını yeterince kapsaması gerekir. Ayrıca MÖ metodunun uygulanabilmesi ve verimli çalışılabilmesi için öznitelik vektörünün yeterince küçük boyutta olması gerekir. Bunun için öznitelik vektörüne boyut indirgeme metotları (Feature selection) uygulanır. Bu metotlardan en çok kullanılanlar: belli metriklere göre öznitelik puanlandırılması ve eşik değer uygulanması (threshold) ile öznitelik seçimi ve bazı ölçüm, optimizasyon (mRmR vb.) yöntemleriyle en uygun öznitelik seti seçme yöntemleridir.

Bu özniteliklerin reel değerleri, ilgili özniteliğin ilgili metinde bulup bulunmama (presence) durumu, ilgili metindeki frekansı (Term Frequency-TF), tüm metinlerdeki frekansı (Document Frequency-DF) ve bu değerlerin belli yaklaşımlarla elde edilen kombinasyonları (TF-IDF) metrikleri kullanılabilir. Bu çalışmada yukarıda bahsedilen metriklerden TF-IDF kullanılarak öznitelik elemesi ve seçimi gerçekleştirilmiştir.

3.3.1 N-Gram modeli

N-gram dil modeli n-1 dereceden bir Markov Modeli sıralamasında bir sonraki elemanı tahmin eden istatistiksel bir dil modelidir. N-gram modeller olasılık, istatistiksel doğal dil işleme, biyolojik gen sırası analizi ve olasılık gibi belli dizilimlerin olasılıklarını inceleyip modelleyen birçok alanda çokça kullanılmaktadır. Daha detaylı anlatmak gerekirse bir n-gram modeli, önceki n elemanlı sıralamanın olma olasılığı bilindiği

takdirde sıradaki olayın olma olasılığını tahmin etmeye çalışır. Bu n-gram modeli doğal dil modellemek için kullanıldığında n-1. sıradan daha önceki kelimeler ile bağımsızlık varsayımı uygulanır ve ilgili kelimenin olma olasılığı sadece kendinden önceki n-1 kelimeye bağlı kılınır. Bu model dilin gerçek yapısını öğrenme problemini, dili yeterince temsil edebilen, büyük miktarda derlemi (corpus) gerekli kılan bir basitliğe indirger.

Çizelge 3.2: Örnek bir cümlede n-gram grupları.

Metin	"okula gitmek için evden çıktı. Ancak başka bir yere gitti."
Unigramlar	'okula', 'gitmek', 'için', 'evden', 'çıktı', 'Ancak', 'başka', 'bir', 'yere', 'gitti'
Bigramlar	'okula gitmek', 'gitmek için', 'için evden', 'evden çıktı', 'Ancak başka', 'başka bir', 'bir yere', 'yere gitti'
Trigramlar	'okula gitmek için', 'gitmek için evden', 'için evden çıktı', , 'Ancak başka bir', 'başka bir yere', 'bir yere gitti'
N-gramlar (n=4)	'okula gitmek için evden', 'gitmek için evden çıktı', 'Ancak başka bir yere', 'başka bir yere gitti'

Doğal dil işlemede n-gramlar, özellikle kelime ve harf sıralamalarının bulunması sürecinde çokça kullanılmaktadır. Kelime n-gramlarından bahsetmek gerekirse; unigram model kendisinden önceki 0 kelime sırasına bağlı iken bigram model kendisinden önceki 1 kelimeye trigram model kendisinden önceki 2 son kelime sırasına bağlıdır (Çizelge 3.2). Konuşma tanıma gibi problemlerde harf ve fonem sıralamalarının tahmininde sıkça kullanılmaktadır. Bu çalışmada n-gram dil modeli kullanılmamış, sadece kelime n-gramları öznitelik olarak kullanılmıştır.

Birçok DDİ çalışmasında bag-of-words metodu kullanılmaktadır. Metinleri sırasız ve gramer bilgisinden yoksun bir şekilde ele alan bag-of-words metodunda, bu şekliyle, yüksek oranda bilgi kaybı kaçınılmazdır. Kelimelerin birçoğu tek tek ele alındıklarında yeterince bilgi içermezlerken, n-gramlar, bileşik kelimeler ve deyimler olarak yan yana geldiklerinde daha yüksek seviyede ve anlamlı bilgiler içerebilmektedir. Bu bağlamda n-gramlar, DA için duygu barındıran kelime sıraları elde etmemizi sağlayan yapılar olarak kullanılabilirler. N-gram'lar, DA için MÖ tekniğinde sıkça kullanılmaktadır. Öncelikle tüm olası n-gram'lar bulunur ve ilgili metin ve veri kümesindeki istatistiksel değerlerine (TF, IDF) göre sıralanıp uygun değerde olanları öznitelik olarak kullanılmaktadır [29] [30].

3.3.2 Olumsuzluk durumları

Doğal dilde olumsuzluk, bazı özel kelime ve eklerin, ilgili kelimelerin veya içinde bulunduğu cümlenin taşıdığı anlamı tersine çevirmesiyle oluşur. İngilizcede olumsuzluk, “not”, “no”, “never”, ”any” gibi kelimeler ve “any-”,”un/in-” gibi öneklerle yapılmaktadır. Türkçede ise olumsuzluk iki farklı şekilde yapılabilir. Bunlardan birincisi; “değil” ve “yok” kelimelerinin ilgili kelime veya kelime gruplarından sonra getirilmesiyle (“güzel değil”), ikincisi de “-me/ma” olumsuzluk eklerinin kullanılmasıyla (“sev-me-di”) yapılmaktadır.

Türkçede ele alınması gereken olumsuzluk bildiren durumlardan biri “-me/ma” olumsuzluk ekleridir. Bu olumsuzluk bilgisinin ortaya çıkarılabilmesi ancak biçimbirimsel çözümlemeyle mümkündür. Bu bilgi yakalandığı takdirde, ilgili olduğu kelimey veya cümley, bir sonraki aşamada işlenmek üzere, özel bir işaretle verilmektedir. Bu işaret “değil” gibi olumsuzluk bildiren kelimelerden birinin ilgili kelime ve kelime grubunun sonuna yerleştirilmesiyle yapılabilmektedir.

Bag-of-words metodu metindeki kelimeleri sıra düzeni ve dil bilgisel kurallarından yoksun ele aldığından, MÖ tabanlı DA’nde, “değil” gibi olumsuzluk bildiren kelimelerin ilgili olduğu kelime ve kelime gruplarını yakalamak mümkün olmamaktadır. Bunun üstesinden gelebilmek için kullanılan n-gramlar veya olumsuzluk durumlarını önişlemlerle işaretleme metotları, başarımı önemli ölçüde arttırmaktadır [3][17]. KDM’ler, kernel fonksiyonlarını kullanarak bu tür birliktelikleri (“güzel değil” gibi) yakalayarak yeni öznitelikler oluşturabildiklerinden, başarımı daha yüksek olabilmektedir. Benzer durum sözlük tabanlı DA için de geçerlidir. Olumsuzluk bildiren kelimelerin hangi kelime ve kelime gruplarıyla ilgili olduğu bilgisi çıkarıldıktan sonra bunların duygusal değerleri belli işlemlere tabi tutularak (işaretinin değiştirilmesi gibi) hesaba katılmaktadır.

4. DENEYSEL ÇALIŞMALAR

Bu bölümde, kullanılan veri kümeleri ve DA için uygulanan metotlar hakkında detaylar verilmektedir.

4.1 Veri Kümeleri

MÖ ve sözlük tabanlı DA metotlarının performanslarını ölçmek ve karşılaştırmak için iki farklı kararkteristikte veri kümesi kullanılmıştır. Birinci veri kümesi, günümüzün en çok kullanılan mikro blog sitesi Twitter'dan elde edilen tweetlerden oluşturulmuştur. Diğer veri kümesi ise "beyazperde.com" adresinden elde edilen sinema filmleriyle ilgili yorumlarından oluşturulmuştur.

4.1.1 Twitter veri kümesi

Twitter veri kümesi imlâ ve dil bilgisi kuralları bakımından oldukça zayıf bir veri kümesidir. Sınırlı metin girişi özelliklerine sahip mobil cihazlarda, sınırlı karakterlerle yazılan tweetler, kısaltmalar, harf eksiklikleri ve kuralsız yapılardan dolayı DDİ için zor metinlerdir. Bunun bir diğer nedeni ise Twitter'ın kullanıcılarına her tweet için 140 karakter kullanma hakkı tanımış olmasıdır. Bu durumdan kaynaklı, kullanıcılar fikirlerini kısaltılmış kelimeler ve işaretlerle anlatmaya çalışmaktadır. DDİ'nin birçok seviyesinde bu tip metinler işlenirken, kurallı ve editör kontrollü metinlere göre daha düşük başarım vermektedirler. Twitter ve film yorumları veri kümelerinin yazım kalitelerinin ölçüsü, Çizelge 4.2'te biçimbirimsel olarak çözümlenemeyen kelime oranları ve benzersiz kelime oranlarıyla gösterilmektedir. Tabloda görüldüğü gibi Twitter veri kümesindeki çözülemeyen kelime oranı ve benzersiz kelime oranı film yorumları veri kümesine göre ciddi oranda yüksektir. Benzersiz kelime sayıları ve veri kümelerinde geçiş sayıları detaylı olarak Şekil 4.1'deki grafikte verilmektedir. Bu grafikte diğer veri kümelerinden daha düzgün yazılmış bir haber metni veri kümesi [31] kullanılmıştır. Haber metni veri kümesi yapılan çalışmada kullanılmamış olup sadece veri kümelerinin düzgünlüğünün karşılaştırılması amacıyla kullanılmıştır. Bu

grafikte her veri kümesinde belli miktarlarda geçen kelime sayıları ve aynı sonuçların veri kümelerinin köklerinin bulunmuş hallerindeki durumları gösterilmektedir. Bu grafikten, haber metinleri veri kümesinin en düzgün yazılmış veri kümesi olduğu ve Twitter veri kümesinin film yorumları veri kümesine göre ne kadar kuralsız ve yazım kalitesinden yoksun olduğunu görülmektedir.

Çizelge 4.1: Twitter veri kümesinde kullanılan alanlar (domainler).

Konu Başlığı	**Alan (Domain)**	**Tweet Sayısı**
Recep Tayyip Erdoğan	Politika	1015
Galatasaray	Spor	1690
Turkcell	Telekominikasyon	775
Mercedes	Otomotiv	200
Arçelik	Elektronik Ev Aletleri	580
Vestel	Elektronik Ev Aletleri	64

Twitter'dan 5 farklı alandan 6 farklı başlıkla ilgili toplanılan tweetlerden 4324 tanesi elle olumlu, olumsuz ve nötr olarak etiketlenmiştir.

Toplanılan bu tweetlerin ilgili oldukları konu başlıkları ve alanlar, Çizelge 4.1'te verilmiştir. Twitter veri kümesindeki tweetler, Twitterdaki 140 karakter sınırlamasından dolayı kısa metinler olup ortalama 14 kelimeden oluşmaktadırlar (Çizelge 4.2).

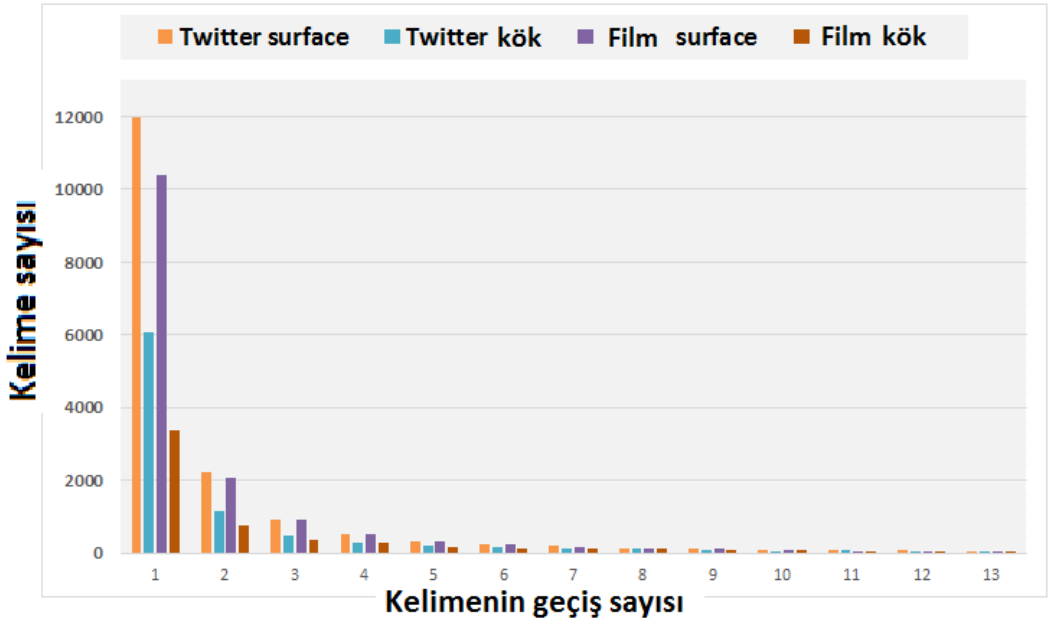
Çizelge 4.2: Twitter ve film yorumları veri kümelerinin özellikleri.

Veri Kümesi	**Olumlu**	**Olumsuz**	**Nötr**	**Toplam**	**Ortalama Kelime Sayısı**	**Çözümlenemeyen Kelime Oranı**	**Benzersiz Kelime Oranı**
Twitter	1677	1301	1346	4324	14	%25	%30
Film Yorumları	13224	7020	-	20244	38	%11	%28
Haber	-	-	-	101000	50	%6	%23

4.1.2 Film yorumları veri kümesi

MÖ ve sözlük tabanlı DA metotlarını karşılaştırabilmek için Twitter veri kümesine göre daha uzun ve dilbilimsel kurallara uygun yazılmış metinlerden oluşan ikinci bir veri kümesi oluşturuldu. Film yorumları tweetlere göre daha kurallı ve imlâ

kuralları gözetilerek yazılmışsa da bu tür yorumlarda oyuncular, yönetmenler ve filmler sıklıkla karşılaştırıldıklarından, birçok duyguyu aynı anda barındırabilme durumları vardır. Bu durum sınıflandırmada başarımı düşürebilmektedir. Örneğin, ilgili filmin yönetmeninin diğer filmleri bolca övüldükten sonra söz konusu filmle ilgili son cümlede kötü yorum yapılmış olabilmektedir. Bu durumda asıl filmle ilgili yorum olumsuz olmasına rağmen genel anlamda yönetmen ve yönetmenin diğer filmeleri övülmüş olduğundan bu yorumu doğru işaretlemek sistemler için zor olmaktadır.



Şekil 4.1: Twitter ve film yorumları veri kümelerinde kelime kök halleri kullanıldığında belli sayılarda geçen kelimelerin sayılarındaki değişim.

Bu veri kümesi, geniş bir yelpazede filmler sunan ve bu filmler hakkında kullanıcıların yorum yapabilmelerine olanak tanıyan "beyazperde.com" adlı sitedeki kullanıcı yorumlarından oluşturuldu. Kullanıcılar film hakkında yorum yaparken filmi beğenip beğenmediklerini belirten ve 1-5 yıldız aralığında yarım yıldızlarla puanlama yapabilmektedirler. Şekil 4.2'te film yorumlarına ve puanlama şekline birkaç örnek bulunmaktadır.

Yorumlar, yazarları tarafından belirlenen yıldız sayılarına göre olumlu-olumsuz olarak işaretlendi. Yüksek başarımlı bir etiketleme yapabilmek için 0.0-2.5 aralığında yıldızla işaretli yorumları olumsuz 4.0-5.0 aralığında yıldızla işaretli yorumları olumlu olarak işaretlendi. Film yorumları veri kümesinin kurallı ve düzgün yazımına bir gösterge olarak biçimbirimsel çözümleyici tarafından çözülemeyen kelime oranı ve benzersiz

Şekil 4.2: Film yorumları veri kümesindeki yorumlar ve puanlandırma şekli.

kelime oranı Çizelge 4.2'da verilmektedir. Görüldüğü üzere film yorumları veri kümesinin her iki değeri de Twitter veri kümesine göre daha düşüktür. Bu da film yorumları veri kümesinin Twitter veri kümesine göre daha kurallı ve düzgün yazıma sahip olduğunu göstermektedir. Benzer biçimde Şekil 4.1'de her iki veri kümesinden eşit kelime uzunluğundaki parçaların (50000 kelime) özellikleri görülmektedir. Bu grafikte verinin normal ve kökleri bulunmuş hallerinde belli sayılarda geçen farklı kelime sayıları verilmiştir. Burada da görülebileceği gibi film yorumları veri kümesindeki düşüş Twitter veri kümesine oranla daha yüksektir. Bu da Twitter veri kümesinin daha bozuk bir yapıya sahip olduğunu ve daha fazla çözülemeyen kelime barındırdığını göstermektedir.

Aynı platformdan benzer şekilde elde edilmiş ve işaretlenmiş başka bir veri kümesi üzerinde Eroğul [8] makine öğrenimi tabanlı, Vural ve diğ. [6] ise sözlük tabanlı yöntemlerle DA çalışmaları yapmışlardır. Aynı veri kümesi üzerinde bu çalışma test edilmek istenmiş ancak ilgili veri kümesine ulaşılamamıştır. Bundan dolayı benzer özelliklere sahip yeni bir veri kümesi oluşturulmuştur. Bu veri kümesi üzerinde elde edilen sonuçların ilgili çalışmaların sonuçları ile karşılaştırması sonuç kısmında verilmektedir.

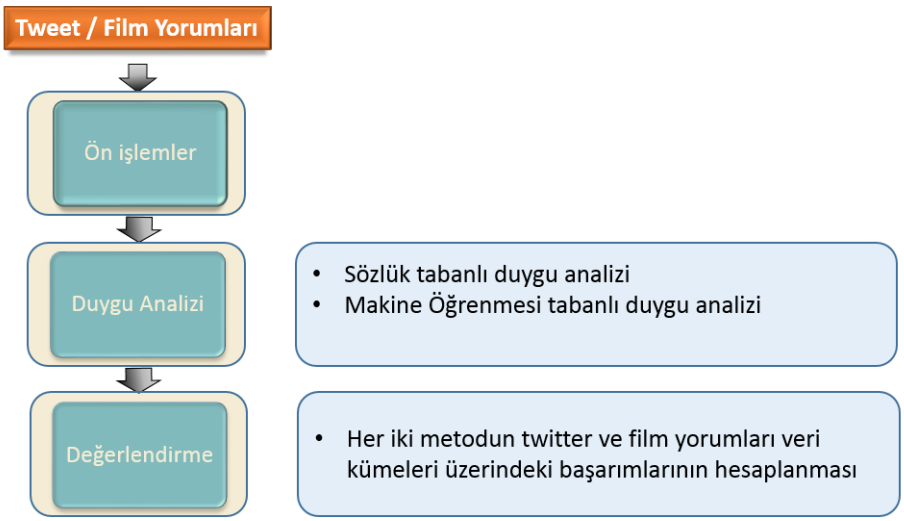
4.2 Kullanılan Metotlar

Uygulanan metotlardan bahsetmeden önce Türkçe'nin özelliklerinden ve bu özelliklerden doğan ön çalışmalardan bahsetmek yerinde olacaktır. Türkçe sondan eklemeli diller grubuna girer. Kelimeler, sonlarına çok sayıda yapım ve çekim eki alarak şekil, biçim ve anlam değiştirebilirler (Çizelge 4.3). Bu anlam değişikliği, kelimelerin taşıdığı duyguyu da değiştirebileceğinden DA için çok kritik bir noktadadır.

Yapılan çalışmalar 3 ana başlıkta toplanabilir: Ön çalışmalar, Sözlük tabanlı DA ve makine öğrenimi tabanlı DA (Şekil 4.3).

4.2.1 Ön çalışmalar

Türkçede DA'nde ilgili kelimelerin duygu durumunu değiştirebilecek birkaç kelime ve ek vardır: olumsuzluk bildiren kelimeler ("değil" (not), "yok" (there is not)), olumsuzluk bildiren ekler (+me/+ma) ve varlık/yokluk ekleri (+lı/+li (with), +sız/+siz (without)). Olumsuzluk bildiren kelimelerin ele alınması sözlük tabanlı DA'nde



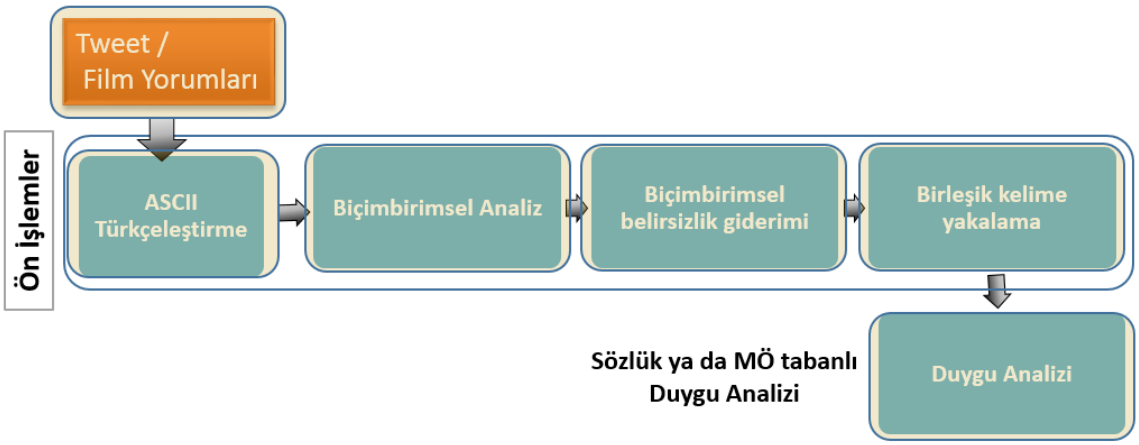
Şekil 4.3: Sistemin genel yapısı.

Çizelge 4.3: Sondan eklemeli bir dil olan Türkçe'nin genel yapısı ve olumsuzluk eki .

Kelime Formları	**İngilizce anlamları**	**Biçimbirimsel Etiket (POS tag)**	**Sözlükteki Duygu Skoru**
iyi	Good	Sıfat	+2
iyileş(mek)	(to) improve	Fiil	+2
iyileştir(mek)	(to) make sb/sth improved	Fiil	+2
iyileştir**me**(mek)	not (to) make sb/sth improved	Fiil	-2
iyileştir**me**yen	the one which does not make sbd/sth improved	Isim	-2

başarımı kayda değer ölçüde arttırdığı daha önceki çalışmalarda belirtilmiştir [2] [5] [6] [7]. Ancak, özellikle Türkçe için olumsuzluk yaratan önemli bir faktör olan varlık/yokluk ekleri ve birleşik kelimelerin kullanılması ilk defa bu çalışmada ele alınmıştır.

Türkçenin sondan eklemeli olmasından kaynaklı özellikleri, veri kümesindeki yazım ve gramer kuralları eksikliği gibi durumların ele alınabilmesi ve biçimbirimsel çözümleyicinin başarılı çalışabilmesi için veri kümesinin ön işlemlerden geçirilmesi önem arz eder (Şekil 4.4). Bu ön işlemler; Türkçe olmayan ASCII karakterlerin uygun Türkçe karakterlerine dönüştürülmesi ve basit yazım hatalarının düzeltilmesi olarak belirtilebilir. Bu işlemler için Zemberek kütüphanesinin [28] ASCII'den



Şekil 4.4: Yapılan ön işlemlerin şeması.

Türkçeleştirme ve biçimbirimsel çözümleyici araçlarından faydalanılmıştır. Bir önceki paragrafta bahsedilen olumsuzluk bildiren durumların ele alınabilmesi, kelime köklerinin ve biçimbirimsel etiketlerin bulunabilmesi için ise biçimbirimsel çözümleme yapılmıştır.

4.2.1.1 Metinlerin temizlenmesi

Veri kümelerinde duygusal olarak anlam ifade etmeyen, metnin kelimelere ayrılması aşamasında sorun çıkaracak olan ve makine öğrenimi metodu için gereksiz yere öznitelik oluşturacak web sitesi ve resim linkleri temizlenmiştir.

4.2.1.2 Normalleştirme

Resmi formatta yazım gerektirmeyen Twitter, Facebook ve SMS gibi platformlarda, bazı kelimeler yazılırken çoklu harf tekrarları yapılabilmektedir. Bu gibi durumlar, genellikle verilmek istenen duygu ve mesaj daha vurgulu verilmek istendiğinde kullanılmaktadır. Bu tür kelimeler (“seviyooorummm”, “çooook”) yakalanıp öncelikle tekrar eden harfleri teke indirmiş daha sonra Zemberek kütüphanesindeki biçimbirimsel çözümleyiciye verilmiştir. Eğer Zemberek kütüphanesi kelimeyi bu haliyle çözümleyebiliyorsa bu haliyle, çözümleyemiyorsa, bu harfler tekrar ikiye (Türkçede her harf en fazla iki tekrarlı bulunabildiğinden) çıkararak ele alınmıştır. Bu şekilde sosyal medyada ve resmi olmayan platformlarda sıkça rastlanan ve gürültü oluşturan bir durumdan kurtulmaya çalışılmıştır.

4.2.1.3 ASCII'den Türkçeleştirme

Türkçede, İngilizcede olmayan, 8 ayrı özel karakter ("ç" , "ş" , "ğ" , "ı" , "ö" , "ü") vardır. Birçok bilgisayar ve mobil cihazda bu Türkçe karakterler bulunmadığından, resmi olmayan yazışmalarda, özellikle sohbet, forum, sosyal medya ve sms gibi platformlarda bu karakterler yerine, bunlara en yakın ASCII karakterler ("c" , "s" , "g" , "i" , "o" , "u") kullanılmaktadır. Bu nedenlerden dolayı veri kümelerindeki birçok metinde Türkçe olmayan karakterler (ASCII) kullanılmıştır. Örneğin ASCII karakterlerle yazılmış "dusurdu" kelimesi, "düşürdü" olarak Türkçe formuna dönüştürülmelidir. ASCII'den Türkçeleştirme işlemini yapabilmek için Zemberek [28] kütüphanesinin Türkçeleştirme modülü kullanılmıştır.

4.2.1.4 İmlâ kontrolü ve düzeltimi

Zemberek kütüphanesi ayrıca Türkçe imlâ kontrolü imkânı sağlamaktatır. Türkçeye uygun olmayan kelimeler için ise yine en yakın kelimeyi önerme özelliği vardır. Zemberek kelime kökünde 3 harf ve eklerinde 2 harf olmak üzere yanlış karakter veya yerleri yanlış karakterleri düzeltme özelliğine sahiptir. Bu özellik sadece 2 harf fark yakınlığında öneri alınabilen, çözümlenemeyen kelimeler için kullanılmıştır.

4.2.1.5 Biçimbirimsel Çözümleme

Biçimbirimsel Çözümleme, cümle içerisindeki her kelimenin kök ve eklerine ayrıştırılması ve görevlerinin belirlemesi sürecidir. Biçimbirimsel çözümlemede kelimelerin kök ve eklerinin çözümlenmesi ile beraber kelimelerin tipi de (isim, fiil, sıfat, zarf, edat gibi.) belirlenir. Biçimbirimsel çözümleme için Oflazer'in Türkçe biçimbirimsel analiz kütüphanesi [9] kullanılmıştır. Buradan kelimelerin ekleri, kökleri ve görevleri (türlerini) elde edilerek gerekli yerlerde kullanmak üzere işlenmiştir. Türkçe, sondan eklemeli bir dil olmasından kaynaklı, biçimbirimsel çözümleme sonucunda, diğer dillere nazaran, daha fazla belirsizlik oluşturur. Bu belirsizlik bir kelimenin birden fazla biçimbirimsel çözümlemesinin olmasından kaynaklanır.

4.2.1.6 Biçimbirimsel Belirsizlik Giderme

Biçimbirimsel belirsizlik giderme, biçimbirimsel çözümleyicinin cümle içerisindeki her kelime için verdiği birçok sonuçtan doğru olanı bulmak olarak ifade edilebilir. Biçimbirimsel çözümleyiciden çıkan birden fazla çözümlemeden en uygunu, en olası olanı seçilmelidir. Bunu gerçeklemek için Sak ve diğ. [10]'nin biçimbirimsel belirsizlik giderici aracı kullanıldı. Bu araç çıktı olarak bir metnin her kelimesinin en olası biçimbirimsel analizini verir.

4.2.1.7 Birleşik kelime çıkarımı

Birleşik kelime çıkarımının amacı, metin içerisinde sıralı ya da sırasız olarak anlam bütünlüğü yaratan kelime bölümlerini yakalamaktır. Birleşik kelimelerin yakalanması, DA açısından da önemlidir. Çünkü ayrı olarak bulunduklarında farklı anlamlar ve duygular barındıran kelimeler bir arada ele alındıklarında daha farklı anlamlar kazanıp daha farklı duygu durumları ifade edebilirler. Örneğin, "kafayı yemek" kelime öbeğindeki kelimeler ayrı ayrı ele alındıklarında olumlu ya da olumsuz bir duygu ifade etmezlerken, bir arada ele alındıklarında olumsuz bir anlam ("psikolojik olarak çökmek") ifade ederler. Ayrıca, "adam olmadı" gibi kelime öbeklerinde olumsuzluk ekinin sadece "ol+mak" kelimesini değil de "adam_ol+mak" birleşik kelimesini etkilemesi gerektiğini anlamak ve ona göre olumsuzluk yaratabilmek için "adam_ol+mak" kelime öbeğinin "adam_ol+eylem" birleşik kelimesi olarak ele alınması önemlidir. Birleşik kelimeler için örnek Çizelge 4.4'de görülebilir.

Bu tür durumların yakalanıp birleştirilmesi ve tek bir kelime olarak ele alınıp o şekilde işlem görmesinin sağlanmasının DA'nde performansı olumlu yönde etkileyeceğini söylemek mümkündür. Olumsuzluk ekinin birleşik kelimenin sadece son kelimesini değil de tüm birleşik kelimeyi olumsuzlaştırması gerektiği açıktır. Birleşik kelimelerin işlenmesi, olumsuzluk durumlarının daha başarılı çalışmasını sağlamaktadır. Birleşik kelimelerin yakalanması için Oflazer'in birleşik kelime çıkarım aracı [11] kullanılmıştır.

Çizelge 4.4: Birleşik kelimeler ve anlam değişimi.

Birleşik Kelime	**Birleştirilmiş Hali**	**Harfi Harfine İngilizce Karşılığı**	**İngilizce Karşılığı**	**Sözlükteki Duygu Skoru**
Kafayı ye-	Kafa_ye+eylem	Eat the head	To get mentally deranged	-3
Adam ol-	Adam_ol+eylem	Be a man	Be a good man	+2
Kafayı çek-	Kafa_çek+eylem	To pull the heads	Consume alcohol	-4
Güzel ol-	Güzel_ol+eylem	Be beautiful	Being beautiful	+3

4.2.2 Metotlar

Biçimbirimsel çözümleme ve diğer önişlem aşamalarından elde edilen kök, gövde ve eklerin gerekli ve etkili bir biçimde kullanacağı yaklaşımlar bu bölümde verilmektedir.

4.2.2.1 Kelime köklerinin kullanılması

Sondan eklemeli bir dil olan Türkçede, teorik olarak, bir kelime sonsuz sayıda ek alabilir ve aynı kökten türeyen çok sayıda kelime oluşturabilir. Bu eklerin her biri, kelimenin anlamını, zamanını, durumunu ve türünü değiştirebilir. Bu yüzden kelime köklerinin kullanılması, özellikle sondan eklemeli diller için önemli bir ön işleme aşamasıdır. Bütün bu kelimelerin barındırdıkları duygu değerleri saptanıp bir sözlükte tutulması ve DA için karşılaştırılarak işlenmesi çok güç olduğundan, kelimelerin köklerinin bulunması ve duygu değerleriyle sözlüğe eklenmesi daha uygundur.

İngilizcede kelimenin anlamı, aldığı sınırlı sayıda ek ile değişmediğinden, “isolate”, “isolated”, “isolation”, ”isolating” kelime çeşitliliği “isolat*” gibi düzenli ifadeler kullanarak tek köke indirgenebilir. İngilizcede “stem” olarak ifade edilen bu kelime kökü (“isolat”) diğer bütün kelime varyasyonlarının yerine gerek makine öğrenimi yaklaşımlarında öznitelik gerekse sözlük tabanlı yaklaşımlarda sözlüğe eklenerek kullanılabilir.

Türkçe için durum daha farklıdır. Türkçe için kelime kökleri bulunurken anlamsal değişiklikler yaratan eklerin atılmamasına ya da atılsa bile bu anlamın saklanabilmesi adına kelimelerin bir şekilde işaretlenmesine dikkat edilmesi gerekir. Bu eklerden en önemlisi, DA’nde etkili olan olumsuzluk bildiren eklerdir. Olumsuzluk ekleri

(+me/+ma) ve varlık/yokluk ekleri (+lı/+li (with), +sız/+siz (without)), bu bağlamda Türkçe için ele alınması gereken eklerdir.

4.2.2.2 Olumsuzluk durumlarının ele alınması

Türkçede olumsuzluk iki farklı şekilde yapılabilir. Bunlardan birincisi, İngilizcedeki "not" kelimesinin karşılığı olan “değil” ve “yok” olumsuzluk bildiren kelimelerin ilgili kelime veya kelime gruplarından sonra getirilmesiyle (“güzel değil” gibi), ikincisi ise ilgili kelimelerde “-me/ma” olumsuzluk eklerinin kullanılmasıyla (“sev-me-di”) yapılmaktadır.

Olumsuzluk bildiren durumlardan biri olan “-me/ma” olumsuzluk eklerinin ortaya çıkarılabilmesi için biçimbirimsel çözümleme gerekmektedir. Bu bilgi yakalandığı takdirde, ilgili kelimenin hemen arkasına “değil” kelimesi eklenerek gerekli bilgi muhafaza edilir. Böylece olumsuzluk bildiren ekler, olumsuzluk bildiren kelimelerden biri haline gelmiş olur ve bu şekilde hesaba katılır.

Olumsuzluk bildiren kelimeleri (“güzel değillerdi”, “onuru yoktu”), biçimbirimsel çözümlemede kök haline getirip (“güzel değil”, “onur yok”) muhafaza edip duygu yoğunluğu hesaplama sırasında ele alınır. Duygu yoğunluğu hesaplama safhasında bu kelimelerin duygu barındıran kelimelerden sonra gelmeleri durumunda ilgili kelimenin barındırdığı duyguyu tersine çevrilmek için sözlükteki duygu puanı -1 ile çarpılarak hesaba katılmaktadır.

4.2.2.3 Varlık/Yokluk eklerinin ele alınması

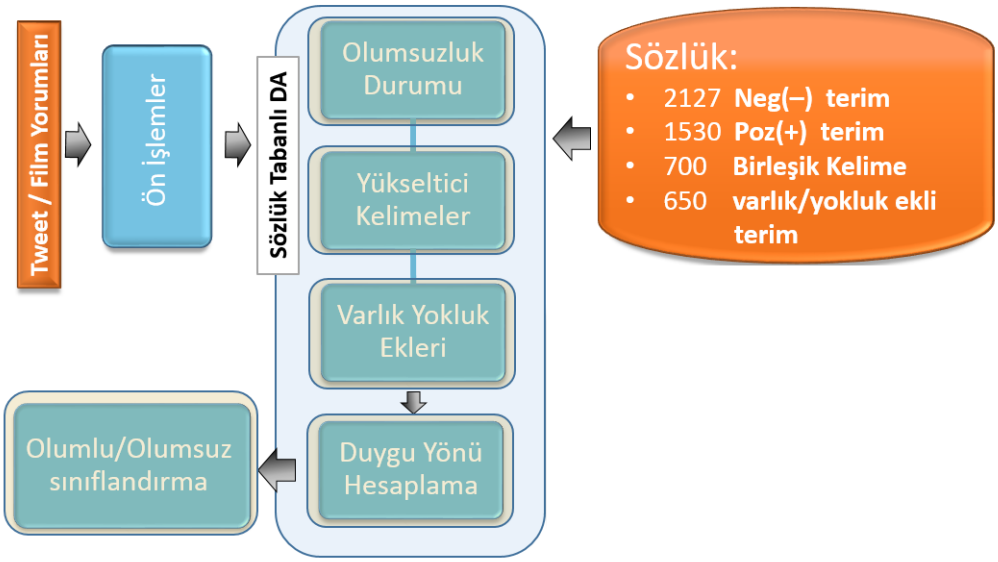
Türkçede ele alınması gereken bir diğer durum ise varlık/yokluk (wit/without) ekleridir (+lı/+li (with),+sız/+siz (without)). İngilizcedeki with/without kelimelerine karşılık gelen bu yapım ekleri isimlerden sıfat yaratırken bu sıfatların olumlu yada olumsuz olmala durumlarını da belirler.

Bu tür ekler barındıran kelimeler biçimbirimsel çözümleme safhasında eklerinden arındırmadan saklanır. Aynı şekilde sözlüğümüzdeki duygu barındıran isimler de olumlu ve olumsuz olacak şekilde bu eklerle beraber saklıdır (“onur+lu (with honor)”, ”onur+suz (without honor)”). Böylece hem sözlük karşılaştırmasında doğru

karşılaştırmayı yapabiliriz hem de makine öğrenimi için daha anlamlı öznitelikler oluşturabiliriz.

4.2.3 Sözlük tabanlı duygu analizi metodu

Sözlük tabanlı DA metodunda sınıflandırma, metindeki kelimelerin daha önceden belirlenmiş, her kelimesinin duygu skorunun olduğu, bir sözlükle karşılaştırılmasına dayanır. Metindeki kelimelerin sözlükteki kelimelerle eşleşmesi gerektiğinden, metindeki tüm sözcüklerin sözlükteki kelime formuna dönüştürülmesi gerekir. İngilizce benzeri dillerde kelimeler fazla ek almadıklarından fazlaca ön çalışma yapılmasına gerek yoktur [12]. Türkçe, sondan eklemeli yapısından dolayı önemli ön çalışmalar gerektirir. Bu ön çalışmalar kelime köklerinin bulunması ve gerekli eklerin işlenmesi olarak belirtilebilir. Bu kısımda sözlük tabanlı DA'nde duygu skorlarının nasıl hesaplandığı detaylı olarak anlatılmaktadır. Sistemin genel görünümü Şekil 4.5'de verilmektedir.



Şekil 4.5: Sözlük tabanlı DA şeması.

Sözlük tabanlı DA metodumuz bütün alanlarda çalışabilecek şekilde tasarlanmıştır. Bu amaçla sözlüğümüz belli bir alana göre değil, genel amaçlı hazırlanmıştır. Sözlük alan bağımsız olmayıp belli bir alana (Finans, sinema, spor, politika) yönelik hazırlansaydı başarım daha yüksek olabilirdi. Bunun nedeni ise birçok kelimenin farklı alanlarda farklı anlamlar taşıyor olmasıdır. Örneğin, "tax","cost","foreign" gibi kelimeler genel olarak olumsuz anlam taşırken, finans alanı için nötr kelimelerdir [32].

Çizelge 4.5: Duygu sözlüğünün içeriği ve kelimelerin duygu değerleri.

Kelimeler	**Duygu Değerleri**
adaletli	2
adaletsiz	-2
adam_et+eylem	2
adam_ol+eylem	2
adapte_et+eylem	2
adapte_ol+eylem	2
adi	-4
adil	2
afalla+eylem	-2
afaroz	-2
afet	-3
affet+eylem	2

Bir metindeki duygunun ortaya çıkarılmasında kullanılacak, her kelimesi taşıdığı duygu yönelimine göre puanlandırılmış bir sözlük oluşturuldu. Bu sözlük; elle Türkçeye çevrilmiş bir İngilizce duygu sözlüğünün [2], Türkçe için ek kelimeler, birleşik kelimeler ve varlık/yokluk ekleri taşıyan kelimeler ile genişletilmesiyle hazırlanmıştır. Kelime köklerine inildiğinden ve kelimeler farklı türde iken farklı anlamlar taşıdıklarından dolayı sözlükteki kelimler ad-sıfat veya fiil olmak üzere iki ayrı etiketle etiketlenmiştir. Sözlükteki her eleman [-5,+5] arasında duygu değerine sahiptir. Cümle içerisindeki kelimelerin işlenişi ve duygu yoğunluğu hesaplanmasında nasıl kullanıldıkları Çizelge 4.5'de örnek cümle üzerinde gösterilmektedir.

Ayrıca önüne geldikleri kelime veya kelime öbeklerinin taşıdığı anlam ve duygunun şiddetini artıran ya da azaltan kelimelerden oluşan bir Booster sözlüğü de oluşturuldu. Bu sözlük; "çok", "en", "az" gibi toplamda 20 kelimeden oluşmaktadır. Metnin duygu değeri hesaplanırken bu yükseltici kelimeler, kendilerinden sonra gelen ve duygu bildiren kelimelerin duygu değerlerine çarpan olarak yükseltici veya düşürücü katkı yaparlar. Cümle içerisinde yükseltici kelimelerin nasıl kullanıldıkları Çizelge 4.6'de örnek cümle üzerinde verilmektedir.

İlgili birçok çalışmada [2] görüldüğü gibi; ünlem işareti ("!"), sonunda kullanıldığı kelimede ya da cümlede, verilmek istenen mesaja ve duyguya vurgu yapmak için

Çizelge 4.6: Yükseltici sözlüğünün içeriği ve kelimelerin çarpım katsayı değerleri.

Yükseltici Kelimeler	**Yükseltme Değeri**
Acayip	2
az	-1
azıcık	-1
aşırı	2
baya	1
cidden	1
en	2
fazla	2
gayet	1
gerçekten	1
kesinlikle	1
yeterince	1
çok	2

kullanılmaktadır. Bu çalışmada da ünlem işareti, ardından geldikleri kelimelerin duygu değerlerini arttıracak zayıf birer yükseltici olarak kullanılmıştır.

Yine ilgili birçok çalışmada [2] [33] gösterildiği ve katkısının belirtildiği gibi; his simgeleri de duygu barındırırlar. Kullanıcılar sosyal medyada kısa yoldan hissiyat durumlarını belirtmek için his simgelerini sıkça kullanılmaktadır.

Her his simgesinin duygu değerinin bulunduğu bir his simgesi sözlüğü oluşturuldu. Bu sözlük en çok kullanılan his simgelerinden oluşmaktadır. Ön işlemler kısmında bu his simgeleri metinlerde aratılarak; olumlu duygu bildirenler “>]” (+1 duygu değerli), olumsuz duygu bildirenler ise “>[” (-1 duygu değerli) şeklinde standart hale getirildi. Bu iki sembol duygu sözlüğüne eklenerek duygu hesaplamasında işleme konulmuştur. His simgelerine birkaç örnek Çizelge 4.7’de gösterilmiştir. Ayrıca, his simgelerinin cümle içerisinde kullanılma şekilleri Çizelge 4.8’de örnek cümle üzerinde gösterilmiştir.

Çizelge 4.7: His simgeleri sözlüğünün içeriği ve simgelerin duygu değerleri.

His Simgeleri	**Duygu Değerleri**
(-:	1
(:	1
(^ ^)	1
(^.^)	1
(^_^)	1
:'(	-1
:'-(	-1
:(	-1
:)	1
:*(	-1
:-(	-1
:-)	1
:-/	-1
:-D	1
:-P	1
;)	1

Sonuç olarak ele alınan bir metin ön işlemlerden geçirilmektedir. Bu aşamada ASCII'den Türkçeleştirme ve imla düzeltimi yapılmaktadır. Daha sonra metinler biçimbirimsel çözümleyiciden geçirilmektedir. Biçimbirimsel çözümleyici ile her kelimenin kökü, türü ve ekleri bulunmaktadır. Buradan çıkan birden fazla sonuçtan en olası olanın seçilebilmesi için bu sonuçlar biçimbirimsel belirsizlik gidericiden geçirilmektedir. Daha sonra birleşik kelimelerin bulunup birleşik duruma getirilmesi için metinler, birleşik kelime çıkarıcıdan geçirilmektedir. Daha sonra biçimbirimsel olarak çözümlenmiş metinde olumsuzluk, varlık/yokluk ekleri yakalanıp saklanmaktadır. Burada önemli olan birleşik kelimelerin ve varlık/yokluk ekli kelimelerin sözlükte bulundukları formlarıyla saklanmasıdır. Son olarak sözlükle karşılaştırılması ve duygu yoğunluğu hesaplanması yapılmaktadır. Bu kısımda; olumsuzluk durumları, yükselticiler, his simgeleri ve kelimelerin duygu değerleri kullanılarak ilgili metnin duygu yoğunluğu hesaplanmaktadır.

Sözlük tabanlı DA metodumuzun genel yapısı Çizelge 4.8'de gösterilmektedir.

Çizelge 4.8: Sözlük tabanlı DA metodunda her modülün, örnek metin üzerinde çalışma şekli.

Tweet	"Galatasaray son macini kazanamadi, maglup oldu ama umutsuz degiliz. Sevgimiz büyük, en iyisi cimbom, Sampiyon cimbom :) "
ASCII'den Türkçeleştirme	"galatasaray son maçını kazanamadı, mağlup oldu ama umutsuz değiliz. Sevgimiz büyük, en iyisi cimbom, Şampiyon cimbom :)"
Biçimbirimsel Çözümleme	. . . maçını maç+Noun+A3sg+P3sg+Acc kazanamadı kazan+VerbˆDB+Verb+Able+Neg+Past+A3sg mağlup mağlup+Adj oldu ol+Verb+Pos+Past+A3sg. . . .
Birleşik kelime çıkarımı	"galatasaray son maç kazan+eylem **mağlup_ol**+eylem ama umutsuz değil sevgi büyük en iyi cimbom şampiyon cimbom >]"
Olumsuzluk	"galatasaray son maç **kazan+eylem değil** mağlup_ol+eylem ama umut**suz** değil sevgi büyük en iyi cimbom şampiyon cimbom >]"
Duygu yoğunluğu hesaplama	"galatasaray son maç kazan+eylem[2][Neg] değil mağlup_ol+eylem[-2] ama umutsuz[-3] [Neg] değil sevgi[3] büyük en iyi[2][Booster] cimbom şampiyon[2] cimbom >][2]"
Duygu değeri ve etiketi	[+14 , -4] -> +10 (olumlu)

4.2.4 MÖ tabanlı duygu analizi metodu

MÖ tabanlı DA metodunda DA'ni bir denetimli sınıflandırma problemi olarak ele alınmıştır. Denetimli sınıflandırma için yeterli boyutta bir eğitim kümesi ve bir test kümesi gerekmektedir. Denetimli sınıflandırıcılar için 10-fold-cross-validation metodunu kullanıldığından, eğitim ve test kümeleri aynıdır ve bütün veri kümesinden oluşmaktadır. 10-fold-cross-validation metodunda veri kümesi 10 eşit parçaya bölünür ve her seferinde bir parça test, kalan 9 parça de eğitim kümesi olarak kullanılır. Daha sonra bu 10 çalışma sonucunda elde edilen başarının ortalaması alınır.

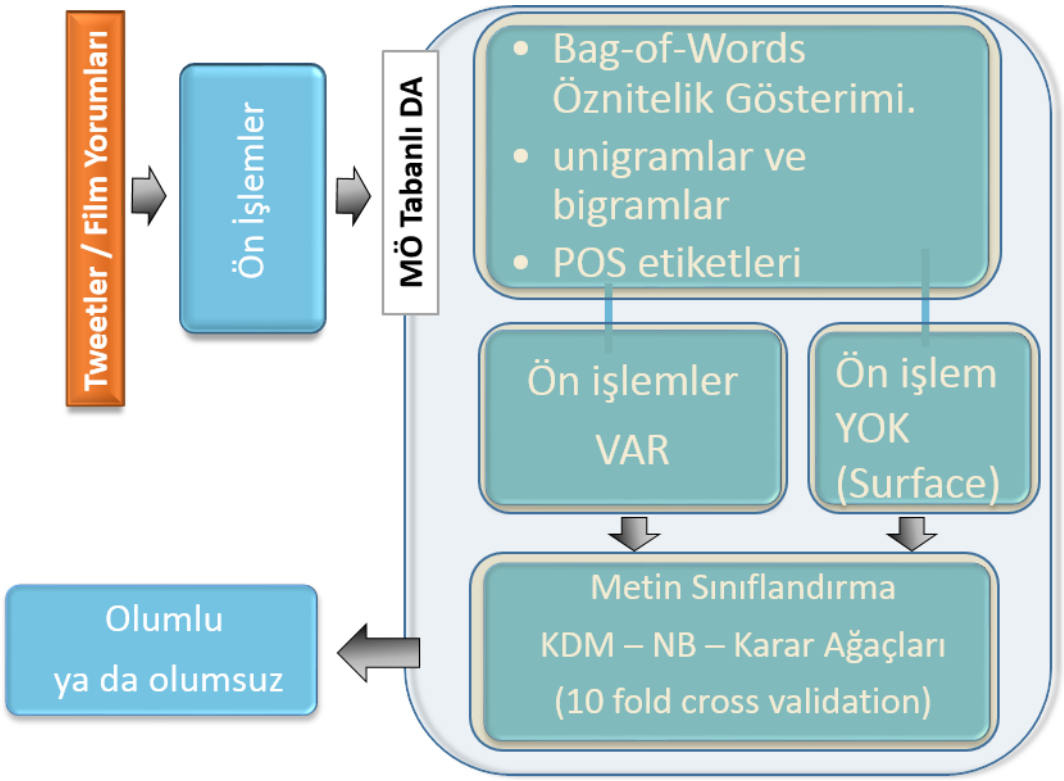
MÖ'ndeki önemli bir nokta öznitelik seçimi ve gösterimidir. Öznitelik olarak uni-gram'lar, bi-gram'lar, POS etiketleri kullanılmıştır. Gerekli boyutta ve etkili

özniteliklerin seçilmesi sınıflandırıcıların performansı açısından çok önemlidir. Öznitelik boyut indirgeme metodları olarak TF/IDF ve mRmR yöntemleri literatürde sıkça kullanılmışlardır. Birçok çalışmada gösterildiği gibi uni-gram ve bi-gram'ların beraber kullanılması, yüksek başarımlar alınabilmesini sağlamaktadır [7] [8].

Bu öz niteliklerin gösterim şekli de diğer önemli bir noktadır. Bu çalışmada öznitelik gösterim şekli olarak bag-of-words metodu kullanılmıştır. Bag-of-words metodu DDİ ve bilgi çıkarımı gibi metin işleyen alanlarda sıkça kullanılan bir öznitelik gösterim şeklidir. Bu modelde her metin kendisini oluşturan kelimelerden oluşan bir yığınla temsil edilir. Bu yığında kelimelerin sırası ve biçimbirimsel yapısı dikkate alınmadığından kelimelerin sıraları ve bu sıralamaların yarattığı önemli bilginin işlenmesi imkansızdır. Bu durum n-gram'lar, deyimler, birleşik kelimeler vb. sıralı durumların yakalanıp tek terime indirgenmesiyle aşılmaya çalışılmaktadır. Bu çalışmada birleşik kelimeler yakalanıp birleştirildiğinden, bu birleşik kelimelerin sıralanması da işlenmiş ve gerekli katkıyı sağlamış olmaktadır. Bu çalışmada kelime kökleri bulunurken fiil olan kelimeler fiil olarak işaretlendiğinden, kelime türleri de öznitelik olarak hesaba katılmış olmaktadır. MÖ tabanlı DA'nin genel şeması Şekil 4.6'de gösterilmektedir.

MÖ'ndeki bir diğer önemli aşama ise sınıflandırıcıların seçimidir. Veri kümesinin karakteristiğine göre en uygun, hızlı ve yüksek başarımlı sınıflandırıcıları seçmek önemli bir konudur. KDM sınıflandırıcısı özniteliklerin birlikteliklerini kullanarak yeni öznitelikler oluşturup kullanma gibi yeteneklere sahiptir. Bu da KDM'nin olumsuzluk, birleşik kelime ve az da olsa cümle içerisindeki kelime sıralamalarını kullanılabilmesini sağlamaktadır.

Bu çalışmada sınıflandırıcı olarak KDM, Naive Bayes ve Karar Ağaçları algoritmaları, bu algoritmalar için ise WEKA [34] aracı kullanılmaktadır. WEKA, öznitelik seçimi ve sınıflandırma için birçok algoritma barındıran bir MÖ aracıdır.



Şekil 4.6: MÖ tabanlı duygu analizi şeması.

5. SONUÇ VE ÖNERİLER

Bu bölümde, uygulanan metotların iki farklı veri kümesi üzerindeki başarımları ve kullanılan her modülün bu başarıma etkisi verilmektedir. Her modülün ilgili metotta ve ilgili veri kümesi üzerindeki katkısı incelenip neden ve sonuç ilişkisi içerisinde tartışılmaktadır.

5.1 Başarımlar

Bu çalışmada kullanılan yaklaşımların ilgili veri kümeleri üzerindeki başarımlarını ölçmek için F1 (F-score) değerini kullanılmıştır. F1 değeri, çağırım (recall) ve hassasiyet (precission) değerleriyle Eşitlik 5.1'de olduğu gibi hesaplanmaktadır.

$$F1 = 2 * [Pressicion * Recall / Pressicion + Recall] \quad \textbf{(5.1)}$$

Sözlük tabanlı DA'de her modülün başarıma etkisinin ölçülmesi için her biri ayrı ayrı inaktif edilerek başarım hesaplamaları tekrarlanmış ve bütün durumları Çizelge 5.1'de gösterilmiştir.

Çizelge 5.1: Sözlük tabanlı DA metodunda her modülün başarıma etkisi.

Sözlük Tabanlı Duygu Analizi Metodu	**Twitter Veri Kümesi**	**Film Yorumları Veri Kümesi**
Modül	**F1**	**F1**
ASCII'den Türkçeye dönüştürme Yok	73.7	74,5
Biçimbirimsel belirsizlik giderme Yok	74.5	77
Olumsuzluk durumlarının ele alınması Yok	72.6	76,5
Yükseltici kelimelerin kullanılması Yok	74.4	77
Bileşik kelime çıkarımı Yok	72.4	78
Varlık/Yokluk eklerinin kullanılması Yok	73.7	77
Bütün modüller Var	**75.2**	**79.5**
Sadece sözlük (Bütün modüller kapalı)	68	71

Makine öğrenmesi tabanlı DA'nde unigram ve bigram özniteliklerinin ektisinin ölçülmesi için bu öznitelikler ayrı ayrı ve beraber kullanılmışlardır. MÖ tabanlı DA metoduyla elde edilen başarımlar Çizelge 5.2'de verilmektedir.

Elde edilen başarımlara göre bütün modüllerin başarıma katkısının olduğu görülmekle beraber en etkili modülün, Twitter veri kümesi için %2.5, film yorumları veri kümesi için %2'lik katkılarla, olumsuzluk durumlarının işlenmesi olduğu görülmektedir. Bu modülü takiben Twitter veri kümesi için %2.5, film yorumları veri kümesi için %1.5'lik katkılarla bileşik kelime çıkarımı işleminin geldiği görülmektedir. Varlık/yokluk eklerinin de birçok DA çalışmasında kullanılan yükseltici kelimelerin kullanımı, ASCII'den Türkçeye çevirme gibi modüller kadar etkili olduğu görülmektedir.

Çizelge 5.2: MÖ tabanlı DA metodunda her öznitelik setinin başarıma etkisi.

Makine Öğrenmesi Methodu	**Twitter veri kümesi**			**Film Yorumları veri kümesi**		
Modül	**KDM**	**NB**	**J48**	**KDM**	**NB**	**J48**
TF-IDF (Unigrams)	84.6	83.7	**81.0**	88.2	87.0	80.0
TF-IDF (Unigrams)-Surface	83.8	82.5	80.4	88.6	88.7	81.9
TF-IDF (Unigram + Bigram)	**85.0**	**84.3**	79.0	**89.5**	**89.5**	**83.0**
TF-IDF (Unigram + Bigram)-Surface	83.7	82.3	77.4	89.0	89.0	82.4

Sözlük tabanlı DA'nin en yüksek başarımları; Twitter veri kümesi için %75.2, film yorumları veri kümesi için %79.5 olarak görülmektedir. Buna karşın Makine Öğrenmesi tabanlı DA'nin en yüksek başarları; Twitter veri kümesi için %85 (KDM), film yorumları veri kümesi için %89.5 (KDM) olarak görülmektedir.

5.2 Tartışma

Görüldüğü gibi iki yöntem de film yorumları veri kümesinde Twitter veri kümesine göre daha başarılılar. Bu sonuçlara bakarak; bu durumun temel nedeni, film yorumları veri kümesinin görece daha düzgün yaılmış yorumlardan oluşması, belli bir alanda (domain) olması ve ilgilenilen konunun sadece hedef sinema filmi olması olarak belirlenebilir. Twitter veri kümesi ise daha bozuk, kuralsız ve kısaltmalarla yazılmış bir metinlerden oluşur. Twitter veri kümesinde toplamda 6 değişik alanla ilgili metinler bulunduğundan, bu veri kümesine alan bağımsız bir veri kümesi diyebiliriz. Her

iki yaklaşımın da Twitter veri kümesinde daha başarısız olmasını bu özelliklerine bağlanabilir.

Sözlük tabanlı DA çalışması denetimsiz bir çalışmadır. Diğer bir deyişle, yüklü miktarda verinin efor sarf edilerek etiketlenmesine ve sistemin eğitilmesine gerek yoktur. MÖ tabanlı DA metoduna göre alan değişimi (Domain Transfer) durumlarına uygundur ve her yeni alandan gelen veriyi sınıflandırmak için o alanla ilgili yüklü miktarda eğitim verisine ihtiyaç duymaz. Twitter verisi çok gürültülü ve zor bir veri olmasına rağmen sözlük tabanlı yaklaşım umut verici sonuçlar vermiştir.

MÖ tabanlı DA metodu her iki veri kümesinde de, diğer birçok çalışmada olduğu gibi, daha iyi sonuçlar vermiştir. Buradan, ilgili veri kümesinden denetimli olarak öğrenen MÖ metodunun, hem uzun (film yorumları) hem de kısa (Twitter) Türkçe veri kümelerinde, daha başarılı sonuçlar verebildiği söylenebilir.

Bu çalışmada etkisi ölçmeye çalışılan önemli iki modülden bileşik kelime çıkarımı, her ne kadar film yorumları veri kümesindeki başarımı düşük olsa da, en etkili ikinci modül olarak ortaya çıkmaktadır. film yorumları veri kümesindeki metinler daha uzun olduklarından bileşik kelimelerin, yakalansa bile, toplam duygu yönelimini değiştirebilecek etkiyi yapamadıkları görülmektedir. Aksine Twitter veri kümesindeki metinler çok kısadırlar ve yakalanan her bileşik kelimenin metnin toplam duygu yönelimini değiştirebilecek etkisi olabilmektedir. Birleşik kelime çıkarımı ve varlık/yokluk ekleri kullanımı gibi gizli bilgilerin ortaya çıkarılıp işlenmesinin umut verici olduğu görülmektedir. Gizli bilginin yanında varolan bilginin cümledeki hangi nesneye yönelik olduğu da çok önemlidir. Daha ileriki çalışmalar için bağlılık analizi yapılarak sadece ilgilendiğimiz nesne ile ilgili kelimelerin dikkate alınması sağlanabilir. Bu şekilde hedefin veya görünümün (aspect) belli olduğu veri kümelerinde sadece ilgili hedefle ilgili özniteliklerin işlenmesi sağlanabilir.

KAYNAKLAR

[1] **Liu, B.** (2010). Sentiment analysis and subjectivity, *Handbook of natural language processing*, ***2***, 627–666.

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

[3] **Gibbs, R.W.** (1986). On the psycholinguistics of sarcasm., *Journal of Experimental Psychology: General*, ***115***(1), 3.

[4] **González-Ibánez, R.**, **Muresan, S. ve Wacholder, N.** (2011). Identifying sarcasm in Twitter: a closer look, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, s.581–586.

[5] **Taboada, M.**, **Brooke, J.**, **Tofiloski, M.**, **Voll, K. ve Stede, M.** (2011). Lexicon-based methods for sentiment analysis, *Computational linguistics*, ***37***(2), 267–307.

[6] **Vural, A.G.**, **Cambazoglu, B.B.**, **Senkul, P. ve Tokgoz, Z.O.**, (2013). A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish, Computer and Information Sciences III, Springer, s.437–445.

[7] **Pang, B.**, **Lee, L. ve Vaithyanathan, S.** (2002). Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, s.79–86.

[8] **Erogul, U.** (2009). Sentiment analysis in turkish, *Middle East Technical University, Ms Thesis, Computer Engineering*.

[9] **Oflazer, K.** (1994). Two-level description of Turkish morphology, *Literary and linguistic computing*, ***9***(2), 137–148.

[10] **Sak, H.**, **Güngör, T. ve Saraçlar, M.** (2007). Morphological Disambiguation of Turkish Text with Perceptron Algorithm, *CICLing 2007*, ciltLNCS 4394, s.107–118.

[11] **Oflazer, K.**, **Say, B.** *ve diğerleri* (2004). Integrating morphology with multi-word expression processing in Turkish, *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, Association for Computational Linguistics, s.64–71.

[12] **Annett, M. ve Kondrak, G.**, (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs, Advances in artificial intelligence, Springer, s.25–35.

[13] **Das, S. ve Chen, M.** (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards, *Proceedings of the Asia Pacific finance association annual conference (APFA)*, cilt 35, Bangkok, Thailand, s. 43.

[14] **Cambria, E. ve Hussain, A.** (2012). *Sentic computing*, Springer.

[15] **Strapparava, C. ve Valitutti, A.** (2004). WordNet Affect: an Affective Extension of WordNet., *LREC*, cilt 4, s.1083–1086.

[16] **Pennebaker, J.W.**, **Mehl, M.R. ve Niederhoffer, K.G.** (2003). Psychological aspects of natural language use: Our words, our selves, *Annual review of psychology*, ***54***(1), 547–577.

[17] **Davidov, D.**, **Tsur, O. ve Rappoport, A.** (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon, *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, s.107–116.

[18] **Socher, R.**, **Perelygin, A.**, **Wu, J.Y.**, **Chuang, J.**, **Manning, C.D.**, **Ng, A.Y. ve Potts, C.** (2013). Recursive deep models for semantic compositionality over a sentiment treebank, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Citeseer, s.1631–1642.

[19] **Bengio, Y.**, **Schwenk, H.**, **Senécal, J.S.**, **Morin, F. ve Gauvain, J.L.**, (2006). Neural probabilistic language models, Innovations in Machine Learning, Springer, s.137–186.

[20] **Jiang, L.**, **Yu, M.**, **Zhou, M.**, **Liu, X. ve Zhao, T.** (2011). Target-dependent twitter sentiment classification, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, s.151–160.

[21] **Turney, P.D.** (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, s.417–424.

[22] **Pang, B. ve Lee, L.** (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, s.271.

[23] **Nguyen, L.T.**, **Wu, P.**, **Chan, W.**, **Peng, W. ve Zhang, Y.** (2012). Predicting collective sentiment dynamics from time-series social media, *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, ACM, s. 6.

[24] **Meral, M. ve Diri, B.** (23-25 nisan 2014). Twitter Üzerinde Duygu Analizi, *SİU 2014(IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı), Trabzon*, Trabzon.

[25] **Simsek, M. ve Ozdemir, S.** (2012). Analysis of the relation between Turkish twitter messages and stock market index, *Application of Information and Communication Technologies (AICT), 2012 6th International Conference on*, IEEE, s.1–4.

[26] **Url-1**, `http://en.wikipedia.org/wiki/Support_vector_machine`, alındığı tarih: 24.04.2014.

[27] **Hsu, C.W.**, **Chang, C.C.**, **Lin, C.J.** *ve diğerleri*, (2003), A practical guide to support vector classification.

[28] **Akın, A.A. ve Akın, M.D.** (2007). Zemberek, an open source NLP framework for Turkic Languages, *Structure*, ***10***.

[29] **Clarkson, P. ve Rosenfeld, R.** (1997). Statistical language modeling using the CMU-cambridge toolkit., *Eurospeech*, cilt 97, s.2707–2710.

[30] **Stolcke, A.** *ve diğerleri* (2002). SRILM-an extensible language modeling toolkit., *INTERSPEECH*.

[31] **Sak, H.**, **Güngör, T. ve Saraçlar, M.**, (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus, Advances in natural language processing, Springer, s.417–427.

[32] **Loughran, T. ve McDonald, B.** (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance*, ***66***(1), 35–65.

[33] **Zhang, L.**, **Ghosh, R.**, **Dekhil, M.**, **Hsu, M. ve Liu, B.** (2011). Combining lexiconbased and learning-based methods for twitter sentiment analysis, *HP Laboratories, Technical Report HPL-2011*, ***89***.

[34] **Hall, M.**, **Frank, E.**, **Holmes, G.**, **Pfahringer, B.**, **Reutemann, P. ve Witten, I.H.** (2009). The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, ***11***(1), 10–18.

ÖZGEÇMİŞ

Ad Soyad: Cumali Türkmenoğlu

Doğum Yeri ve Tarihi: Bozova - 22.03.1985

Adres:

E-Posta: turkmenogluc@itu.edu.tr

Lisans: Kocaeli Üniversitesi

Y. Lisans: İstanbul Teknik Üniversitesi

Mesleki Deneyim ve Ödüller:

Yayın ve Patent Listesi:

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- **Türkmenoğlu, C.**, Tantuğ, A. C., (2014). Sentiment Analysis in Turkish Media, *International Conference on Machine Learning (ICML 2014),Workshop on Issues of Sentiment Discovery and Opinion Mining, Beijing, 2014*