DEVELOPING A TURKISH SENTIMENT LEXICON USING TONE DISTRIBUTIONS

TON DAĞILIMLARI KULLANAN TÜRÇE DUYGU SÖZLÜĞÜ GELİŞTİRİLMESİ

MUAZZEZ ŞULE KARAŞLAR

ASSOC. PROF. DR. BURKAY GENÇ
Supervisor

Submitted to
Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

August 2022

ABSTRACT

DEVELOPING A TURKISH SENTIMENT LEXICON USING TONE DISTRIBUTIONS

Muazzez Şule KARAŞLAR
Master of Science, Computer Engineering
Supervisor: Assoc. Prof. Dr. Burkay GENÇ
August 2022, 65 pages

With the developing technology and increasing use of the internet, many sources of data have been exposed to researchers. Analysis and extraction of meaningful information from this data is a research topic under the field of natural language processing. Sentiment analysis which is a sub-field of NLP evaluates the content of data with respect to the opinion it conveys as one of positive or negative. Most sentiment analysis research is done using one of two approaches: lexicon based and machine learning based. Lexicon based approach needs a dictionary of positive and negative words which are used to evaluate a text. Although there are abundance of studies in English, the same can not be claimed for Turkish. Therefore, in our study, we focus on constructing a comprehensive and accurate Turkish sentiment lexicon.

In this paper, we aim to develop a Turkish sentiment lexicon with a novel methodology: using statistical tone density functions computed using a very large document corpus obtained from mainstream Turkish news agencies. In this way, for the first time in the literature, a Turkish sentiment lexicon is created by using this method. The lexicon not only assigns tone values instead of boolean polarities, but also provides sharper tones which is usually not possible with other approaches in the literature. We evaluate the performance of this lexicon in comparison with similar lexicons in the literature. Results show that the constructed sentiment lexicon in this study achieves a comparable performance and poses many potential improvement possibilities.

Keywords: Sentiment analysis, Natural Language Processing, lexicon, polarity, statistical distribution.

ÖZET

TON DAĞILIMLARI KULLANAN TÜRÇE DUYGU SÖZLÜĞÜ GELİŞTİRİLMESİ

Muazzez Şule KARAŞLAR
Yüksek Lisans, Bilgisayar Mühendisliği
Danışman: Assoc. Prof. Dr. Burkay GENÇ
Ağustos 2022, 65 sayfa

Gelişen teknoloji ve internet kullanımının artmasıyla birlikte birçok veri kaynağı araştırmacıların kullanımına açılmıştır. Bu verilerden anlamlı bilgilerin çıkarılması ve analiz edilmesi Doğal Dil İşleme (DDİ) alanında bir araştırma konusudur. DDİ'nin bir alt alanı olan duygu analizi, verilerin içeriğini, verdiği görüşe göre olumlu veya olumsuz olarak değerlendirir. Çoğu duygu analizi araştırması, iki yaklaşımdan biri kullanılarak yapılır: sözlük tabanlı ve makine öğrenimi tabanlı. Sözlük tabanlı yaklaşım, daha sonra bir metni değerlendirmek için kullanılan olumlu ve olumsuz kelimelerden oluşan bir sözlüğe ihtiyaç duyar. İngilizce'de çok sayıda çalışma olmasına rağmen Türkçe için aynı şeyi söylemek pek de mümkün değildir. Bu nedenle, bu çalışmada kapsamlı ve doğru bir Türkçe duygu sözlüğü oluşturmak amaçlanmıştır.

Bu çalışma kapsamında, ana akım Türk haber ajanslarından elde edilen kapsamlı bir döküman bütünü kullanılarak ve hesaplanan istatistiksel ton yoğunluğu fonksiyonunu kullanarak yeni bir metodolojiyle Türkçe duygu sözlüğü geliştirmeyi amaçlıyoruz. Bu sayede literatürde ilk kez bu yöntem kullanılarak Türkçe duygu sözlüğü geliştirilmiştir. Bu sözlük, kelimelere yalnızca ikili polariteler yerine ton değerleri atamakla kalmaz, aynı zamanda literatürdeki diğer yaklaşımlarla genellikle mümkün olmayan daha keskin ton değerleri elde edilmesini sağlar. Bu çalışmada, elde ettiğimiz sözlüğün performansını literatürdeki benzer sözlüklerle karşılaştırmalı olarak değerlendiriyoruz. Sonuçlar oluşturulan duygu sözlüğünün karşılaştırılabilir bir performansa ulaştığını ve birçok potansiyel iyileştirme olanağı sunduğunu göstermektedir.

Anahtar Kelimeler: Duygu analizi, Doğal Dil İşleme, duygu sözlüğü, polarite, istatiksel dağılım.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Burkay Genç, who guided me in this study for 3 years.

I would like to thank Fatih Sağlam for his support and ideas.

Last but not the least, I would like to thank my parents and my husband Mustafa Danyıldız for their support without them it would not be possible.

CONTENTS

Page

ABSTRACT i
ÖZET iii
ACKNOWLEDGEMENTS v
CONTENTS vi
TABLES viii
FIGURES ix
ABBREVIATIONS x
1. INTRODUCTION 1
1.1. Scope Of The Thesis 2
1.2. Contributions 2
1.3. Organization 2
2. BACKGROUND OVERVIEW 4
2.1. Natural Language Processing and Sentiment Analysis 4
2.1.1. Natural Language Processing 4
2.1.2. Sentiment Analysis 7
2.1.3. Applications for Sentiment Analysis 8
2.1.4. Sentiment Analysis Levels 8
2.1.4.1. Document Level 9
2.1.4.2. Sentence Level 10
2.1.4.3. Aspect/Feature Level 10
2.1.5. Sentiment Analysis Methods 11
2.1.5.1. Machine Learning 11
2.1.5.2. Lexicon based 13
2.1.5.3. Hybrid Approach 14
2.2. Sentiment Lexicon 15
2.2.1. Manual construction 15
2.2.2. Automatic Construction 16

2.3. The GDELT Project 17
2.3.1. GDELT Databases 18
2.3.1.1. GDELT 2.0 Event Database 18
2.3.1.2. GDELT 2.0 Global Knowledge Graph (GKG) 19
2.4. The Zemberek Library 20
2.5. Confusion Matrix 21
3. RELATED WORK 23
4. PROPOSED METHOD 27
4.1. Dataset 27
4.2. Data Preprocessing 29
4.2.1. Tokenization 30
4.2.2. Morphology 30
4.2.3. Remove Stop Words 31
4.3. Creating Sentiment Lexicon 31
4.3.1. Weighted Average Approach 31
4.3.2. Tone Density based Approach 32
4.3.3. Sentiment Tone Results 34
4.3.3.1. Positive Terms 34
4.3.3.2. Negative Terms 35
4.3.3.3. Close to Neutral Terms 36
4.4. MLTC (Manually Labeled Turkish Corpus) 38
5. EXPERIMENTAL RESULTS 39
5.1. Weighted Average Approach 39
5.2. Tone Density based Approach 40
5.3. Merged Method 41
5.4. SWNetTR++ 41
5.5. Test Result Evaluation 43
6. CONCLUSION 44

TABLES

Page

Table 2.1 Basic format of an entry in the Event Database 19
Table 2.2 Basic format of an entry in the Event Database 20
Table 2.3 Confusion Matrix 21
Table 4.1 Tone vector of ”velet” 32
Table 4.2 Test Set Distribution 39
Table 5.1 Confusion Matrix for weighted average method 40
Table 5.2 Precision, recall, F1-Score, accuracy values for weighted average method 40
Table 5.3 Confusion Matrix for Tone Density based Method 41
Table 5.4 Precision, recall, F1-Score, accuracy values for Tone Density based Method 41
Table 5.5 Confusion Matrix for Merged Method 42
Table 5.6 Precision, recall, F1-Score, accuracy, values for Merged Method 42
Table 5.7 Confusion Matrix for SWNetTR++ 42
Table 5.8 Precision, recall, F1-Score, accuracy, values for SWNetTR++ 42
Table 5.9 Evaluation Results for 300 News 44

FIGURES

Page

Figure 2.1 Sentiment Analysis Levels 9
Figure 4.1 Published news 28
Figure 4.2 GDELT Document Tone Values 29
Figure 4.3 Probability of density function curve 33
Figure 4.4 The relationship between PDF and CDF 33
Figure 4.5 Sentiment Tone of ”mutlu” 34
Figure 4.6 Sentiment Tone of ”sevgi” 34
Figure 4.7 Sentiment Tone of ”beraberlik” 35
Figure 4.8 Sentiment Tone of ”başarı” 35
Figure 4.9 Sentiment Tone of ”hırsızlık” 36
Figure 4.10 Sentiment Tone of ”ölüm” 36
Figure 4.11 Sentiment Tone of ”tehikeli” 36
Figure 4.12 Sentiment Tone of ”uyuşturucu” 36
Figure 4.13 Sentiment Tone of ”deniz” 37
Figure 4.14 Sentiment Tone of ”sağlık” 37
Figure 4.15 Sentiment Tone of ”yağmur” 37
Figure 4.16 Sentiment Tone of ”sıcaklık” 37
Figure 4.17 Sentiment Tone of ”engelli” 38

ABBREVIATIONS

NLP	:	**N**atural **L**anguage **P**rocessing
ML	:	**M**achine **L**earning
BoW	:	**B**ag **o**f **W**ords
SVM	:	**S**upport **V**ector **M**achine
POS	:	**P**art **o**f **S**peech
GDELT	:	**G**lobal **D**atabase **o**f **E**vents **L**anguage **a**nd **T**one
AI	:	**A**rtificial **I**ntelligence
TF	:	**T**erm **F**requency
IDF	:	**I**nverse-**D**ocument-**F**requency
RF	:	**R**andom **F**orest
NB	:	**N**aive **B**ayes
KNN	:	**K**- **N**earest **N**eighbor
GKG	:	**G**lobal **K**nowledge **G**raph
URL	:	**U**niform **R**esource **L**ocator
MLTC	:	**M**anually **L**abeled **T**urkish **C**orpus

1. INTRODUCTION

Today, with the widespread use of social media and the internet, users have access to large number of interpretable thoughts through the online environment. Due to this rapid increase in data in the electronic environment, the need to automatically analyze the data and extract meaningful information from the data is increasing.

Sentiment analysis, also known as Opinion Mining, is the interpretation of a subjective language element, such as speech or writing, on a particular topic. Most of the data is available in text form and using this analysis method, words and phrases can be classified as positive, negative, and neutral.

Sentiment analysis which can be roughly divided into two which are machine learning and lexicon based approach. In machine learning approach, linguistic features, machine learning models and algorithms are used [1], [2] . On the other hand, lexicon-based approach uses sentiment lexicon for identifying the sentiment tones of words and texts.

For analyzing text content by using lexicon-based approach, existence of the languages own resources is very important and necessary. Compared to other languages including Turkish, English has very rich language resources and libraries such as SentiWordNet [3], OpinionFinder [4], and WordNet-Affect [5]. For this purpose, translation approach from English to target language is mostly used methodology in the literature. However, the use of language libraries obtained with this approach may prevent accurate sentiment analysis because of the different language structures.

To this end, this study aims to develop a comprehensive and accurate Turkish sentiment lexicon using an automated pipeline. Main stream news media sources have been used from the GDELT database to build this sentiment lexicon. However, unlike the literature, lexicon is developed by a novel method: using statistical tone density functions computed using a very large document corpus obtained from mainstream Turkish news agencies.

1.1. Scope Of The Thesis

This thesis mainly focuses on the development of a comprehensive, correctly polarized and accessible Turkish Sentiment Lexicon. Thus, it is aimed to perform sentiment analysis of Turkish content with a higher accuracy. For analyzing text content, existence of the languages own resources is very important and necessary. In many other languages, including Turkish, language libraries and resources are not yet fully developed. In the literature, the translation approach from English language sources is used in the creation of the sentiment lexicon. However, the loss of meaning and emotion arising from the systematic of translation is inevitable, which can negatively affect the accuracy of the sentiment analysis. For this, it is aimed to develop a Turkish sentiment lexicon with a novel approach by using Turkish sources.

1.2. Contributions

In this study, we aim to develop a sentiment lexicon which can be used as a resource in many NLP studies. Turkish sentiment lexicon studies are still underdeveloped and it is hard to find a mature, widely accepted, open source sentiment lexicon for general purpose. We aim to fill this gap. We believe, the theories and practices developed in this thesis may also be used in different languages for developing sentiment lexicons in those languages. Unlike the literature, lexicon is developed by a novel method: using statistical tone density functions computed using a very large document corpus obtained from mainstream Turkish news agencies. In this way, for the first time in the literature, a Turkish sentiment lexicon is created by using our method and this study poses many potential improvement possibilities.

1.3. Organization

The organization of the thesis is as follows:

- Chapter 1 presents our motivation, contributions and the scope of the thesis.

- Chapter 2 provides an overview of the background of our thesis. NLP, sentiment analysis, sentiment lexicon development approaches are discussed. It explains the GDELT database which is used in this study.
- Chapter 3 gives a related works in the literature.
- Chapter 4 introduces our methodology including dataset by developing Turkish sentiment lexicon in detail. Samples of sentiment tone values in lexicon are shown.
- Chapter 5 gives the experimental results. Performance of the lexicons which are developed with different approaches are evaluated and also, comparison of these lexicons are done.
- Chapter 6 states the summary of the thesis and possible future directions.

2. BACKGROUND OVERVIEW

2.1. Natural Language Processing and Sentiment Analysis

2.1.1. Natural Language Processing

Language is the greatest need of social life for communication between people. The fact that over 7000 languages are currently spoken in the world clearly shows the important of this concept. In the computer science, there are two different languages. These are programming languages and natural languages, which are machine languages. Languages that humans use are called natural languages. In today's world, the connection between machine and human is getting closer with increase of social media, blogs, content and text data.

Natural Language Processing is the field of computer science and AI that produces solutions to maximize human-computer interaction by giving computers the ability to understand texts and spoken words similarly to humans. While doing this, it basically uses statistical, machine learning and deep learning models and combines them to create enhanced methodologies. By the help of NLP, computers can process human language in text or speech format and extract the meaning from the data including emotion or intent in it.

The most fundamental characteristic of natural language is that it is one of the most natural and important tools to share feelings and thoughts with other people and form a nation. Therefore, the terms, language rules and structure for each society are different from each other. As a result, the methods, approaches, and especially the tools developed in the field of natural language remain mostly language-specific, making it necessary for researchers to conduct language-specific studies according to the rules and structure of each language.

There are two basic inputs to the process of natural language processing. These are text and audio elements. In academic studies, the focus is on analysis when the input is text and on the response of the system when the input is audio. In both way, system produces form of a structured output.

Applications of Natural Language Processing is explained [6] in detailed.

- Search Autocorrect and Autocomplete
- Language Translator
- Sentiment Analysis
- Social Media Monitoring
- Chatbots
- Survey Analysis
- Targeted Advertising
- Hiring and Recruitment
- Voice Assistants
- Grammar Checkers
- Email Filtering

Search Autocomplete and Autocorrect

Search auto-completion and auto-correction are applications of NLP. Many people use them daily and even expect them when they search for something. Search auto-complete predicts what the user is looking for, so the prediction can be clicked instead of typing further. This helps answer questions faster and find the right information. It also reduces the likelihood that the user will become apathetic and abandon the search.

It's also easy to make a typo and not notice it. Search engines no longer just use keywords to guide users to search results. They analyze users' intentions when searching for information using NLP. Thus, the search engine corrects errors using NLP and still finds relevant results. Search auto-completion and auto-correction help us find the right results very efficiently.

Virtual Assistants and Chatbots

Today, NLP-based virtual assistants and chatbots are used to answer questions automatically. Chatbots are designed to help the user and answer any question. Similarly, virtual assistants use NLP and speech recognition to understand the user's voice commands and respond accordingly. Most chatbots and virtual assistants follow predefined rules when answering questions. Together with powerful AI, they can learn something and respond appropriately by interacting with the user.

Email Filtering

One of the most basic applications of NLP is email filtering. Emails are divided into different sections depending on their content. To avoid cluttering our inbox, spam emails are filtered for specific words and phrases and routed to another section. This is done with the help of NLP's text classification technique. The process of classifying texts according to predefined categories is called text classification.

Machine Translation

Machine translation is one of the first applications of natural language processing. This application area, which is the focus of NLP, attempts to make sense of words and word sequences. In machine translation, a text is automatically converted from one language to another without losing its meaning.

In the past, machine translation systems were dictionary-based and rule-based systems. The success rate was not very high. With the help of development in the field of neural networks, big data acquisition and powerful machines, machine translation has become more and more successful in converting texts from one language to another.

Social Media Analytics

The use of social media is increasing day by day, and many people share their thoughts about politics, a certain product, or a topic. By analyzing these thoughts, useful information can be gained about the person's likes and dislikes. However, with a large number of users, it can be difficult to interpret them. This is where natural language processing and sentiment analysis help us. Companies use various NLP techniques to analyze posts on social media and learn what customers think about their products. Companies also use social media monitoring to understand what problems their customers have when using their products.

Text Summarization

With a large amount of data, it is difficult to review all the data and create a summary. Natural language processing can be used to extract the most relevant information and summarize the text. This enables the simplification of large amounts of data such as scientific articles, news or legal documents, etc. With NLP, data can be summarized in two different ways: using keywords or by creating new paraphrases based on meanings and inferences.

2.1.2. Sentiment Analysis

Sentiment analysis is a sub-field of natural language processing studies. Sentiment analysis, also known as opinion mining, can be used to interpret subjective elements of language in verbal or written texts on a specific topic and determine the emotional meaning of communications.

People usually want to know different experiences, feelings and opinions before acting and evaluate them in the decision-making process. This action, which used to take place in mutual conversations, is realized through social media platforms, which are becoming more and more widespread with the development of technology. On such platforms, people take into account the comments and ratings of other users when choosing hotels for vacation, buying clothes, and even ordering dinner. With sentiment analysis, it is now possible to determine the meaning, emotions, and behaviors of texts such as comments and ratings on social media platforms.

2.1.3. Applications for Sentiment Analysis

Sentiment analysis can be used for various purposes in many different fields by analyzing the emotions in the content of large amounts of text. Such analysis provides extremely important data that many companies, institutions, and organizations can benefit from in determining their strategies. A detailed information of applications of sentiment analysis are explained in [7]. We can list some of them below.

- Determination of customers' attitudes towards a brand or product, their satisfaction or dissatisfaction
- Social media monitoring
- Evaluation of customer complaints
- Development of suitable dialogue systems according to the emotional states of the users
- Determination of people's attitudes and feelings about a particular topic
- Politicians can measure how the public reacts to their campaigns during election periods.
- Listen to voice of the customer (VoC)
- Listen to voice of the employee

2.1.4. Sentiment Analysis Levels

Sentiment analysis can be done at different levels: document level, sentence level or aspect/feature level. These levels are mentioned in detail (Bharat,2017,p.493-494) [8]. Figure 2.1 is shown in this article.

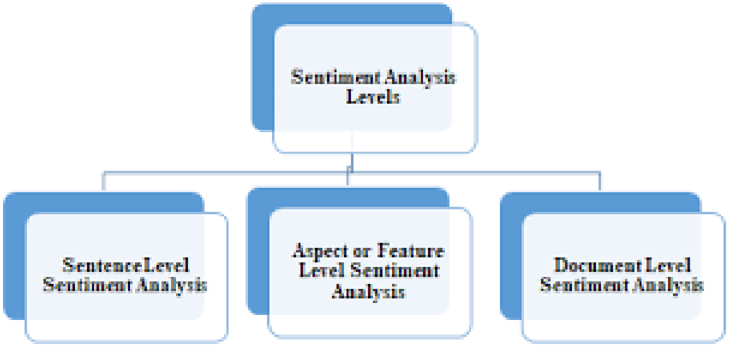

Figure 2.1 Sentiment Analysis Levels

2.1.4.1. Document Level

Sentiment analysis and opinion mining initially originated as document-level classification studies. In document-level sentiment analysis, the entire document is considered as a single thought without going into details, and a classification is made according to the state of expression of positive or negative sentiments.

Since document-level sentiment analysis yields only a single result, it is not suitable for comparing multiple cases. This method mostly preferred in cases where the content target is a single object, for example, in the analysis of online product reviews.

As an example of document level sentiment analysis, Balahur [9] classified the documents according to a range of psychological emotions such as happy, unhappy, angry, and guilty, rather than categorizing them just as positive and negative. They used Appraisal Theory in their studies which in psychology is a field of study that analyzes the process of forming beliefs about assets and events. In their study, ISEAR database has been used which has information about people's reactions to various events and situations, such as their attitudes and behaviors. Balahur and others compared their proposed methods with machine learning based methods and found that their method was more successful.

2.1.4.2. Sentence Level

Sentiment analysis at the sentence level is not much different from sentiment analysis at the document level.In this classification method, after considering whether each sentence is subjective or objective, if the sentence is subjective, a classification is made according to whether the sentence expresses positive or negative emotion.

Sentiment classification at the sentence level provides a more detailed view than sentiment analysis at the document level. At the same time, document-level analysis techniques can be applied to sentences. This classification method does not cope with some sentence structures. For example, interrogative sentences, : ”Is the weather very nice?” or allusive sentences: ”You know best of all”.

Comments people make about a product, topic, or person can be not only positive or negative, but also about the characteristics of the product, topic, or person. In this situation, sentiment analysis at the document and sentence level is not sufficient.

2.1.4.3. Aspect/Feature Level

At the aspect level, sentiment analysis aims to analyze the object in all its aspects, not just positively or negatively. In document and sentence level classification, the comments do not need to be elaborated, but at the aspect level, the comments are elaborated to determine the direction of emotions in relation to specific features of a given object.

Hu and Liu’s feature-based work with the product evaluation dataset in 2004 is the first example of these studies [10]. In this study, Hu and Liu applied a word-based method using a dictionary of emotion terms. In the word-based method, they first determined a certain number of emotion terms (seed) to create the dictionary of emotion terms. Then, they expanded these terms with the antonym and synonym functions of WordNet to create the dictionary of emotion terms. For feature extraction, they proposed a method called ”frequent names”.

Frequent names method shows that in all the comments of the dataset, the words with high frequency are usually the words that express the features of the product. When many product reviews and datasets on this topic are examined, it is found that this assumption is correct. However, not all names that are frequently mentioned in the comments are related to product features. Some common noun-type words may be words other than product features. These words should be identified and separated from words related to product features. This method was later used by many researchers for feature extraction.

In Qiu's approach [11], aspect words and their polarities are analyzed using grammatical features. After determining the target and scope of emotion expression, correct feature extraction is the most important part. While sentiment polarity is difficult to identify, feature extraction is often an even more difficult task. Akba et al. [2] explained in detail the importance of feature extraction metrics in machine learning approaches for sentiment analysis.

2.1.5. Sentiment Analysis Methods

Sentiment polarity detection is the main task in sentiment analysis and opinion mining and is performed in sentiment analysis studies at all levels. The methods used to detect sentiment polarity divided into two main categories. These categories are machine learning based methods and lexicon based [12] methods.

Generally, machine learning based methods are preferred in studies related to a specific domain, while lexicon based methods are preferred in studies related to scalability as a result of the analysis performed in the text.

2.1.5.1. Machine Learning

Machine learning is often used in the classification of emotions. Therefore, various classification algorithms are used. Mostly Support Vector Machine, Naive Bayes, Decision trees and Maximum Entropy classification algorithms are used. In methods based on machine learning, the system needs predetermined training data. Machine learning approaches divided into two categories as supervised and unsupervised.

When supervised learning algorithms are to be used, the texts are needed to be labeled. On the other hand, unsupervised learning algorithms passes the labeling step. Al-Hadhrami et al. [13] used supervised and unsupervised algorithms in their studies and they compared SVM and RF, one of the supervised machine learning algorithms, and K-means clustering algorithms from the unsupervised machine learning algorithms, which are used for sentiment analysis on tweets in English. The properties of unigram and bigram approaches are used for feature selection. As a result of the study, it was seen that SVM outperformed other approaches.

The most important factor affecting the success rate of machine learning methods is the recognition and use of effective features for sentiment classification in the creation of the feature vector.

Davidov and his friends [14] classified emotions in Twitter messages in their study. In creating the feature vector in their studies using the supervised machine learning method, they also used features such as smiley characters and hashtags on web platforms and Twitter that do not appear in other texts, apart from the features commonly used in text mining. They found that using these unique features of Twitter increased success.

The following describes commonly used feature selection methods for creating feature vectors in sentiment analysis studies.

Bag of Words

Bag of Words is the most commonly used feature in text classification based on whether words occur in texts. It considers not only the individual occurrence of words in the text, but also whether they occur in pairs or in groups of n in a series. If the words are treated individually, it is the uni-gram model, if they are considered in groups of two in a row, it is the bi-gram model, and if they are treated in groups of n in a row, it is the n-gram model. These model is an application of vector space model.

TF-IDF (Term Frequency - Inverse Document Frequency)

If a word occurs frequently in only one type of document and very rarely in other types of documents, it has high distinctiveness. However, if a word is frequently used in all documents and not only in a specific genre, it means that its distinctiveness is low. TF-IDF is a method of weighting terms calculated based on the frequency of a word in the document, the total number of documents containing that word, and the number of all documents.

Part of Speech

Word types (noun, adjective, verb, etc.) are an important feature used in natural language processing and word processing studies. Word types are also important in studies of sentiment analysis. In particular, words such as adjectives, verbs, nouns, and adverbs are of great importance. The word types of words in the text are also used to create the feature vector.

2.1.5.2. Lexicon based

This category uses methods based on syntactic analysis of sentences using natural language processing methods and tools. No labeled training data is required as in supervised machine learning methods. Sentences are analyzed using natural language processing tools and methods, and semantic inferences are made by determining the emotional terms in the sentences. A dictionary of emotional terms is usually used to detect emotional expressions in sentences.

Lexicon-based methods aim to analyze emotions using dictionaries consisting of words or phrases that express emotions. Lexicon-based methods can use existing dictionaries to assign sentiment values to terms, or a new lexicon can be created [15]. Lexicon-based methods can be applied mainly to specific patterned texts such as blogs, forums, and product reviews because of high scalability.

Lexicon-based methods are studied under two main headings: corpus-based approaches and dictionary-based approaches. In the dictionary-based approach, the set of manually created sentiment terms is used to find synonyms and antonyms of words using dictionaries such as most used lexicon WordNet. The set of sentiment words is expanded and the search process is terminated if no new word is found [16] . In the corpus-based approach, sentiment lexicon is generated using statistical or semantic methods.

Ding et al. [17] developed a sentiment analysis system that automatically rates tourist attractions based on comments received on Ly.com, a Chinese travel booking platform. The sentiment analysis system was built considering the Chinese sentiment lexicon.

From the collected data, it selected the ratings of the ten best tourist attractions in Wuhan according to the ranking results of the website. The top ten tourist attractions were evaluated using the calculation methods, and at the end of the study, it was found that the ranking results is similar to the ranking of Ly.com.

2.1.5.3. Hybrid Approach

The hybrid approach is a combination of machine learning algorithms and dictionary-based approaches. Rumeli et al. [18] used users' product reviews and ratings on the online shopping platform Hepsiburada.com as a data set and developed a sentiment analysis model that combines machine learning algorithms and dictionary-based approaches. In the first step of the study, a computation was performed by developing a model based on the sentiment value of each word in the sentence in the sentiment lexicon. Then, based on the polarity values of the texts, the machine learning algorithms NB, RF, SVM and KNN were trained to perform sentiment analysis. As a result of the study, the sentiment analysis was performed with 73% accuracy without human intervention.

Another study of using hybrid approach is belongs to Ersahin et al. [19]. They use a sentiment lexicon that is extended with a synonyms lexicon. As ML approach, they use three different algorithms which are NB, SVM and J48 as classification algorithms.

The hybrid approach combines these two approaches by generating a new lexicon-based value according to the feature generation algorithm and feeds it to the machine learning classifiers as one of the features. The studies have shown that this new approach provides a 7% improvement.

2.2. Sentiment Lexicon

The sentiment lexicon is a kind of dictionary that numerically records the emotion aspect (negative/neutral/positive) and emotion evaluation of the terms in its content. Sentiment lexicon can have phrases besides individual words. Sentiment lexicon generation has three main approaches [20] :

- Manual construction
- Automatic construction

2.2.1. Manual construction

Manually constructed sentiment lexicons are created by human annotators. The most costly way to create a sentiment dictionary is to do it manually. Although the size of the dictionary is relatively small, its accuracy is higher than other approaches.

Tuerkmenoğlu and Tantuğ [21] focused their study on comparison of dictionary-based and machine-learning-based sentiment analysis and perform which one has the better result in Turkish. Within the scope of the study, they created Turkish sentiment lexicon by manually translating the dictionary of the tool used in English called SentiStrength [22] into Turkish. In the first phase, they created a Turkish emotion dictionary with a total of 3657 words by adding negativity expressions and other words they thought necessary to the emotion dictionary with 2547 words.

Today, there are also studies to develop a sentiment lexicon, using crowdsourcing over the Internet, with a manual but broad-based participation process. Generally, data is collected from individuals to answer questions prepared or games developed for this purpose, and a lexcion is created.

An example of crowdsourcing approach is study of Mohammad and Turney [23] determined the sentiment polarities of the terms in the dataset they created from various dictionaries using Amazon's Mechanical Turk crowdsourcing platform and the feedback they received from online participants. As a result of the study, which has a systematic question-answer system and provides solutions for effective question selection and elimination of incorrect answers, the English emotion dictionary called EmoLex was developed, consisting of a total of 10170 terms. EmoLex which describes Emotion Lexicon is a list of English words and their associations with eight basic emotions : anger, fear, anticipation, trust, surprise, sadness, joy, and disgust and two sentiments as negative and positive.

The literature indicates that this approach is used to develop a sentiment dictionary that is narrow enough to meet the needs in cases where a sentiment dictionary is not available or cannot be acquired. To make the results more robust, more than one annotators can be assigned to the task and the agreement between the annotators can be done. As a result, there is no algorithm or calculation in this construction methodology.

2.2.2. Automatic Construction

In automatic creation of sentiment lexicons mostly dictionary based and corpus based approaches have been used. In contrast to manually construction, it gives lower accuracy but in case of time and cost parameters, it has more advantages. In dictionary based approach, one of the most used method is using a set of starting seed words with known sentiment alignment and then expand that set using an existing lexical resource. Another way is translation mechanism.

In translation mechanism, a dictionary of emotions is obtained by translating an existing comprehensive dictionary of emotions into the target language. This system is often used in languages other than English where there is no sentiment dictionary and language resources are limited.

Yengi [24] conducted a study based on dictionary based method and implemented recommendation systems using values from sentiment analysis instead of user ratings by using Amazon and Movie Lens data set. In this study, the importance of personalizing recommendation systems with big data analysis methods has been revealed, and the contribution of these systems to the basic methods was reported and the results were analyzed.

The results of the analysis show that the growth of data volume and personalizing this system increases the success rates. However, at the same time, the negative effect of the increase in the volume of data on the error value is also observed.

2.3. The GDELT Project

GDELT is a public database created by Kalev H. Leetaru. It contains current news and publications in many languages and is still under development. The main purpose of the GDELT system is to scan and compile all open sources of information in the world and to encode and present the compiled data in a computational form. This platform monitors the global news media from all regions of the world in more than 100 languages in online and offline formats.

The GDELT project was first released in 2013 in version 1.0. It was further developed with the GDELT 2.0 version in 2015. Although the GDELT project was released in 2013, it has incorporated analysis of news from around the world from 1979 to the present. The archives of major agencies such as AfricaNews, Agence France Presse, Associated Press were used in the background of GDELT, whose news history dates back to 1979.

In the GDELT project, news is not stored as direct news text, but in a format that includes the news URL and processed metadata. GDELT collects news every 15 minutes and translates these news into English in real time for 65 languages.

Each news text is archived in the GDELT databases by using NLP mechanisms to code it by theme, sentiment and tone, location, and entity (organizations and people). This process is repeated every 15 minutes.GDELT's data collection and processing capabilities generate trillions of data points every year thanks to Google Cloud's algorithm-based technology.

The GDELT project offers researchers a variety of ways to use the datasets it contains. With version 1.0 of GDELT, researchers have been able to download GDELT in ".csv" format from file servers, and since 2015, with version 2.0, the datasets have been offered through Google's cloud computing service, Google BigQuery.

2.3.1. GDELT Databases

In the GDELT project, in addition to the various datasets, the two main datasets, the Event Database and the Global Knowledge Graph (GKG), attract the attention of researchers.

2.3.1.1. GDELT 2.0 Event Database

The Event Database collects data from many different news articles and records events. It extracts information about who did what to whom and how many news articles reported it. For each record, it stores actor1, actor2, event, and digitized information about that event. The Event dataset contains records from today to 1979. Structurally, each event has one record row and 61 columns.

In table x you can see the selected columns of the table in the event database. The Event Database Codebook [25] provides a general overview of the fields available in the Event

Database table and their detailed descriptions.

Actor1Code	Actor1Name	Actor2Code	Actor2Name	EventCode	GoldsteinScale	NumMentions	NumSources	NumArticles	AvgTone	DATEADDED	SOURCEURL
EDUEDU	SCHOOL	null	null	120	-4.0	20	2	20	-2.55699408950183	20160506084500	http://www.iha.com.tr/haber-lise-ogrencisi-parkta-olu-bulundu-556954/

Table 2.1 Basic format of an entry in the Event Database

2.3.1.2. GDELT 2.0 Global Knowledge Graph (GKG)

GKG stores news articles by extracting detailed information about all the people, numbers, themes, places, and emotions mentioned in an article. Structurally, each event has one record row and 27 columns. This method was developed based on the idea that contextual analysis is necessary when evaluating tone.

Table x shows an excerpt from the GKG for a Pentagon news article about the statement 'underage fighters' in Syria. For better readability, the excerpt has been reduced to the main columns and the variable content truncated after about 70 characters. For an overview of all the fields of GKG and how they are coded, see the GKG Codebook [26].

DATE	DocumentIdentifier	SOURCECOMMONNAME	V2COUNTS	V2THEMES	V2LOCATIONS
20170228050000	http://aa.com.tr/tr/dunya/ pentagondan-suriyede-kucuk-yastaki-savascilar-aciklamasi/760517	aa.com.tr	null	ALLIANCE;TERROR; WB_2467_TERRORISM; WB_2433_CONFLICT_AND_VIOLENCE; WB_2432	1#Iraq#IZ#IZ##33#44 #IZ#2627;4#Aleppo, ?Alab, Syria#SY#SY09#26319#36.20

V2PERSONS	V2ORGANIZATIONS	V2TONE	ALLNAMES
States Rakka,677	Us Center Forces,68;United States,671; United States,2675;United States	-1.00574712643678,2.29885057471264, 3.30459770114943,5.60344827586207	Forces Commanded,81;Syria North,323; The Photo,428;Syria Democratic,550/

Table 2.2 Basic format of an entry in the Event Database

GDELT analyzes the content of the news using advanced language resources, but does not check the validity of the news text. We tried to maximize the validity of the news we received through GDELT using news from 3 major news agencies (AA, DHA, İHA). Hard-to-confirm news from local and lesser-known sources were ignored.

2.4. The Zemberek Library

The NLP Zemberek library can be used to perform basic NLP operations such as spell checking, tokenization, root finding, morphological parsing, word suggestion, word formation, conversion of words written with ASCII characters, and syllable retrieval [27].

In this study, the processes of parsing Turkish news texts into words (tokenization) and linguistic operations (morphology) were applied by the help of Java-based Zemberek library developed by the Akın brothers.

2.5. Confusion Matrix

The confusion matrix is a table often used to describe the number of correct and incorrect predictions of our model and the performance of the model by including the known real values in the dataset and the prediction values of the classification model. An example of an confusion matrix table is shown in Table 2.3.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2.3 Confusion Matrix

The entries in the confusion matrix can be defined as follows:

True Positives (TP): The true value and the predicted value are positive.

True Negatives (TN): The true value and the predicted value are negative.

False Positives (FP): The true value is negative but the predicted value is positive.

False Negatives (FN): The true value is positive but the predicted value is negative.

There are some ratios calculated from the confusion matrix. These terminologies showing the relationship between TP, TN, FP and FN are listed below.

Accuracy: It is the ratio of predictions made correctly in the system to all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: A state that indicates success in a positively predicted situation.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity): It indicates how well positive situations are predicted.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Specificity: It indicates how well negative situations are predicted.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

F Score: It is the harmonic mean of the Precision and Recall metrics.It is a measure of how well the classifier is performing and is often used to compare classifiers.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

3. RELATED WORK

The lexicon-based approach finds the sentiment polarity of a document by using the sentiment tone of each word in a given document. There are several well- known sentiment lexicons and most of them in English such as SentiWordNet, SenticNet [28]. As seen in the literature, most known sentiment lexicons are in English and there is a lack in other languages sentiment lexicon including Turkish. Because of the this lack, first Turkish sentiment lexicon SentiTurkNet is created by Dehkharghani [29].

While developing SentiTurkNet, they used Turkish WordNet and manually labelled synsets as positive, negative and neutral and ex tracted features for synsets. Then, extracted features are used in machine learning classifiers for developing sentiment lexicon by mapping features with polarity values. They concluded that translation method for developing lexicon is not a best methodology because of the terms in one language can not have a equivalent meaning or sentiment because of the language structure and culture is different.

While this work is extensible for other languages with corresponding WordNets, this also means a limitation of its scope. The Turkish WordNet is not public and the size of 15,000 synsets is much smaller than the English SentiWordNet with 117,000 synsets. It shows the lack of Turkish resources and some studies which depend on this data are limited by size.

In his thesis, Eroğul studied the dataset of movie ratings using a machine learning based method [30]. In his study, he used a Bag of Words (BoW) model for the feature vector and features related to the word type. SVM was preferred as the classification algorithm. Success rates ranging from 72% to 85% were achieved in evaluation and accuracy is performed with different parameters.

The most known lexicon is SentiWordNet which is described in detail [3] which is the result of the automatic annotation of all the synsets of WordNet where each synset has three numerical value that shows how positive, negative, or objective of the terms contained in the synset are. The method used to develop SENTIWORDNET is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification.

In another popular publicly available resource SenticNet [28] which uses common sense reasoning techniques with a model for categorising emotions and an ontology for describing human emotions. In their approach lexicon is labelled using pleasure, attention, sensitivity and aptitude.

There are lots of study which use SentiWordNet by developing sentiment lexicon in their own language. Ucan [31] uses this lexicon for developing Turkish Sentiment Dictionary by using translation approach from English to Turkish. He used SVM (Support Vector Machine) method and movie and hotel reviews as labeled data for evaluating performance of the created lexicon. This lexicon has named as SWNetTR.

To enhance the SWNetTR lexicon, Fatih and Burkay [32] conducted a study. Their study uses GDELT Database to reach Turkish news pages on the web for enriching the lexicon from 27K to 37K words. For determining the performance of this lexicon, they tested the lexicon on domain independent news texts. With this study SWNetTR lexicon expanded with around 10000 unique words and this new lexicon was named as SWNetTR-PLUS and this thesis led to our study.

Another sentiment lexicon which uses Turkish news as a study of Kaya, Fidan, and Toroslu [33] examined Turkish political news in their study on sentiment analysis. Texts such as product reviews are thought-intensive and topic-oriented, while political news is less thought-intensive.

Moreover, political news is not as topic-dependent as product reviews. Kaya et al. used word-based methods in addition to machine learning-based methods. They created a list of sentiment lexicon for political news. In their tests and measurements, they achieved success rates between 66% and 77%.

For developing SWNetTR-PLUS, Fatih and his friends [34] conducted a new study, they used Turkish synonym and antonym word pairs and extended this lexicon by almost 33 percent to obtain SWNetTR++, a Turkish sentiment lexicon with 49K words. They followed a different way in their studies to expand the lexicon by using graph structure by representing words as nodes, and edges as synonym–antonym relations between words.

In recent studies, star ratings of available text documents, such as different product categories on online platforms like amazon, yelp.com, or IMDB, were used to create domain-specific sentiment lexicons. One of the studies that uses this rating data to create sentiment lexicons is [35], which performs a SentiDraw method using the star ratings of the reviews and the probability distribution of these star ratings over the words to compute the sentiment scores.

Sentiment lexicons can be divided into domain-independent and domain-specific lexicons. A word can be interpreted in many ways based on its domain. It is extremely important to demonstrate each word under its domain, because meanings differ greatly according to domain. To this end, Dong and friends [36] propose a novel hierarchical monitoring topic model to develop a domain-dependent sentiment lexicon.

In their sentiment construction model, they consider the effects of topics or domains on the sentiment polarities of words and called as topic-adaptive sentiment lexicon model. This model assumes that each word can have more than one sentiment polarity for different topics. They model hidden topics and emotions jointly, where each word is generated by a conditional polynomial distribution over topic and sentiment pairs.

John et al. [37] (2019) conducted a lexicon-based sentiment analysis study on tweet data with positive, negative, and neutral ratings from sentiment140.com. In the study, sentiment classification was performed using three different lexicon, which are SentiWordNet, Hybrid Lexicon and Hybrid Lexicon followed by Sentiment Adjustment Factors named as ”H+S ADF”.

Classification processes in terms of polarity of the term in the sentiment lexicon may be different for domain specific words, therefore the contextual polarity of the text should be considered. SentiWordNet which is a general purpose lexicon, has accuracy of 79.80% in classification. On the other hand, in domain specific words this classification ratio decrease to the value of 68.20%. When the hybrid lexicon has applied and after sentiment adjustment factors like considering emoticon, modifiers and negation added, the accuracy of this lexicon is increased approximately 74.80%. The results show that the application of the Hybrid Lexicon Classification and Sentiment Adjustment Factors increase the accuracy of the sentiment analysis.

Yurtalan et al. [38] (2019) conducted a lexcion-based sentiment analysis study for Turkish tweets. Their sentiment lexicon includes 1181 data item with positive and negative word roots, part of speech (POS) tags, and polarity values. Author claims that literature has the some Turkish sentiment lexicon with low accuracy because translation method from source language into Turkish can lose the meaning because of the morphological structure of the Turkish language. In this scope, he developed a sentiment lexicon by using POS tags of words as noun, verb, pronoun, adjective, adverb, conjunction etc. and these tags are used for determining word groups. While determining the sentiment polarity of the text, he considers the word groups and POS tags. As a result, study shows that considering the word groups in lexicon based sentiment analysis improves the performance.

4. PROPOSED METHOD

In this section, dataset and sentiment lexicon creation methodology will be explained in detail.

4.1. Dataset

This study uses GDELT which have been explored by various authors for sentiment analysis in academic research. GDELT is a comprehensive database of web news from all over the world in over 100 languages which provides open access to metadata. We used GDELT's Global Knowledge Graph (GKG) database which consist of news articles where entities, themes, locations and tone are coded. It computes thousands of sentiment scores for every article.

While assigning polarity and tone values which uses translation approach. First, all non-English news articles are machine translated into English and then sentiment analysis has been done in English.

All analyzes performed in this study are based on metadata from the database compiled by the GDELT project. When querying the GDELT database, only the message text metadata and the URL of the message text accessed by GDELT are provided. Since this data is not sufficient to perform a comprehensive analysis of the language used, we accessed each URL individually to confirm that the text was still accessible and to import the text into our local database so that it could be processed.

At this stage, we used the newspaper library [39] to clean the news text from HTML tags and other content unrelated to the news on the website. In total, we pulled the URL and metadata information of 162653 news texts published between 2017 and 2018 from the GDELT database.

We found that 158984 of these URLs were still accessible and 3669 of them were inaccessible, and we excluded the inaccessible news data from the study. We took care to distribute the texts evenly over the 2-year interval so that they did not focus on a particular time or event.

We selected three major news agencies (Anadolu Agency, Demiroeren News Agency, and İhlas News Agency) as news sources. As can be seen in Figure 4.1, İhlas News Agency published more individual news stories than the other agencies during this period.

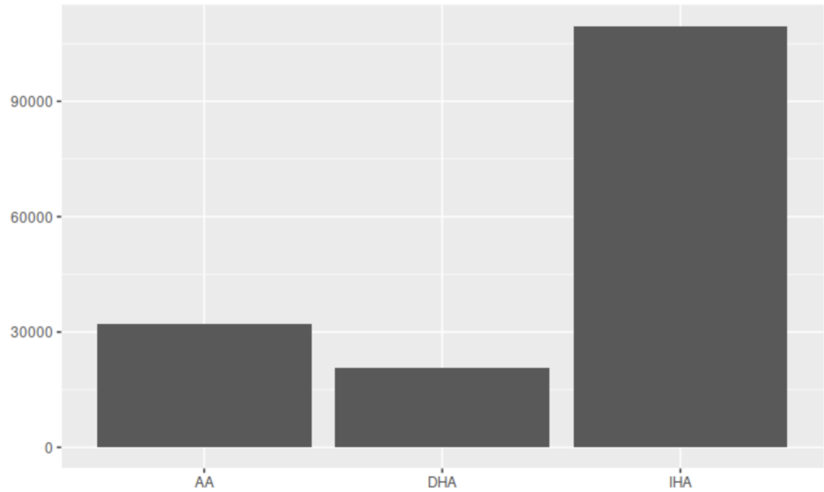

Figure 4.1 Published news

The tone and polarity values from the GDELT database metadata are an important component of our study. The tone value is calculated by translating the text into English by the automatic algorithms used by the GDELT project to measure the emotional intensity of a news text. Although GDELT says that this value can range from -100 to +100, it has been shown that the tone values in the data we obtained were mostly between -10 and +10.

Figure 4.2 shows the distribution of tone values determined by GDELT for the messages we retrieved. Another important point is that the messages with a tone value close to zero are filtered during data extraction from the GDELT database. The reason for this is that monotonous and routine news texts with a very low tone value should not be included in the study database. In this way, we wanted to bring out the emotional tone in the news language as much as possible.

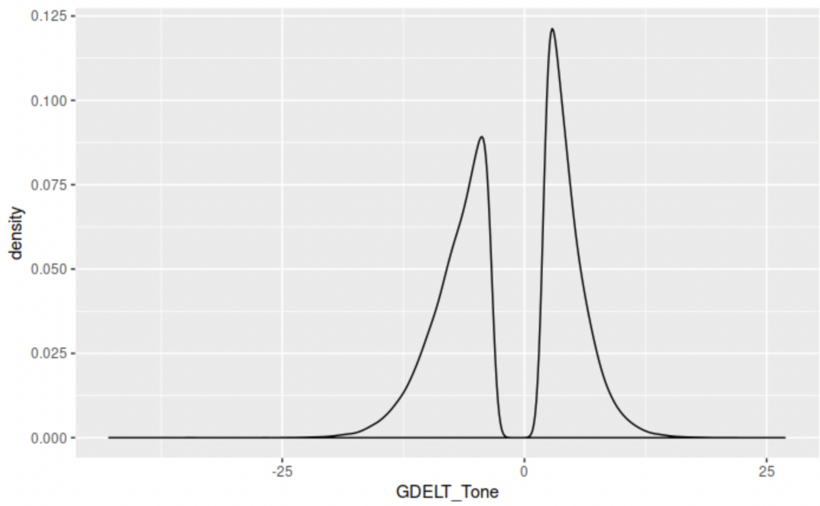

Figure 4.2 GDELT Document Tone Values

Another point that can be clearly seen in the figure is that the negative message tone is more scattered and the positive message tone is more distributed. While the average tone value for the messages we receive in our database with a negative tone is calculated as -7.0904386, this average is calculated as 4.5163524 for the messages with a positive tone value.

There is also a difference in standard deviations of 3.0166378 and 2.2142682. However, the number of negative and positive messages in the database shows a balanced distribution: 77483 positive messages are compared to 81501 negative messages in the database.

Since the tone values offered by GDELT are given in a very wide range and there are almost no values at the ends of the range, we normalized the tone values in the database and reduced them to the range [-1,+1].

4.2. Data Preprocessing

The first operation on any dataset is data preprocessing. Data preprocessing is the work done to make the raw and scattered data ready for analysis. Different steps of text preprocessing are tokenization, noise removal, lowercasing, stopwords removal, stemming, and lemmatization.

In our study, we implemented the step of removal noisy data. The newspaper library [39] was used to clean the news text from HTML tags and other content unrelated to the news on the website as described in the dataset title. The definitions of the next steps such as tokenization, linguistic operations(morphology), and removing stop words and their usage in the study are described in the following sections. In some of these steps, the Zemberek library was used.

4.2.1. Tokenization

The Stanford NLP Group [40] describes tokenization like that: ”Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation”.

When parsing the news texts into words, the ”TurkishTokenizer” class from the Zemberek library was used. Thus, we parsed the news into words. This process also removes expressions such as spaces, new lines, punctuation, Roman numerals, numbers, percentages, time, date, URL, email, HashTag, Mention, Emoji, etc. from the text.

4.2.2. Morphology

Morphology is the study of the internal structure of words. Morphology is concerned with how the components (roots, stems, prefixes, suffixes, etc.) of a word are arranged or changed to produce different meanings. Zemberek library provides morphological analysis functionality.

During the linguistic process, all words are first converted to lowercase and the text is normalized. The ”TurkishMorphology” class of the Zemberek library returns details about the linguistic structure of the given text. The tokenized words were analyzed and decomposed into their suffixes and roots using the methods of ”TurkishMorphology” class. The type of each root form (adjective, noun, verb, etc.) was labelled. This process is called as POS

(Part-of-Speech). Although these steps have been applied to the words in news texts in preprocessing data, they were not used in our study. It is planned as future work.

4.2.3. Remove Stop Words

The Stanford NLP Group [41] describes stopwords like that: ”Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words”.

In the list of stop words, we find words that we use frequently in daily life, but which do not express the meaning of the text, or which are not very familiar to us in Turkish.

4.3. Creating Sentiment Lexicon

In scope of this work, sentiment tones and polarity of each word was calculated using two approach. One of the methods is using weighted average which is often used method in literature by finding sentiment tone, other one is based on tone distribution of terms in documents which is a novel methodology and the lexicon obtained this way, provides sharper tones which is usually not possible with other approaches in the literature.

Applying these methodologies two main metric are used.

- Frequency of words in news text
- The tone value assigned by GDELT to the news text

4.3.1. Weighted Average Approach

In weighted average approach, firstly tone vectors of terms are found and then using the equation (6) weighted average is calculated.

$$S_{\mathrm{T}} = \frac{\sum_{i=i}^{n}(d_i.f_i)}{\sum_{i=i}^{n} f_i} \tag{6}$$

where: n = number of documents containing term

d_i = sentiment tone of the document

f_i = the number of times the term appears in the document

S_T = sentiment tone of the term

Document	Document Tone	Term Frequency
DOC-1	-11,801	1
DOC-2	9,871	1
DOC-3	5.319	2

Table 4.1 Tone vector of ”velet”

The tone vector of the word ”velet” is shown in Table 4.1. The word “velet” has appeared in 3 different new documents, 4 times in total, and 2 of these news have positive sentiment tone and 1 of these has negative tone. With the mathematical expression in Equation 6, the sentiment tone value of the word ”velet” is calculated as -0.185. The document tones shown in Table 4.1 were normalized to be between -1 and 1, and thus the sentiment tone of the term was found between -1 and 1.

4.3.2. Tone Density based Approach

In tone density based approach, sentiment tone is found by a novel methodology: using statistical tone distributions based on the density functions computed over the documents. This approach aims to find sharper sentiment tone for the word by calculating difference of positive and negative areas under the density function.

Total Area under the curve in probability of density function is 1 and shown in the Figure 4.3.

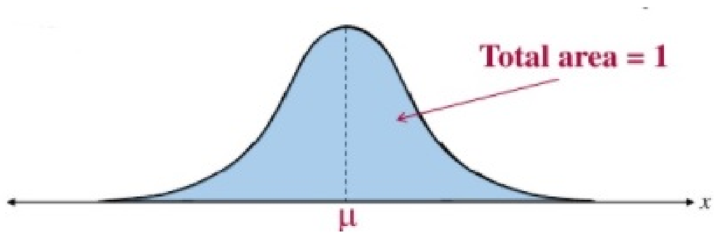

Figure 4.3 Probability of density function curve

In this plot of density function in Figure 4.4, calculation of negative and positive areas are based on the statistical approach described in [42].

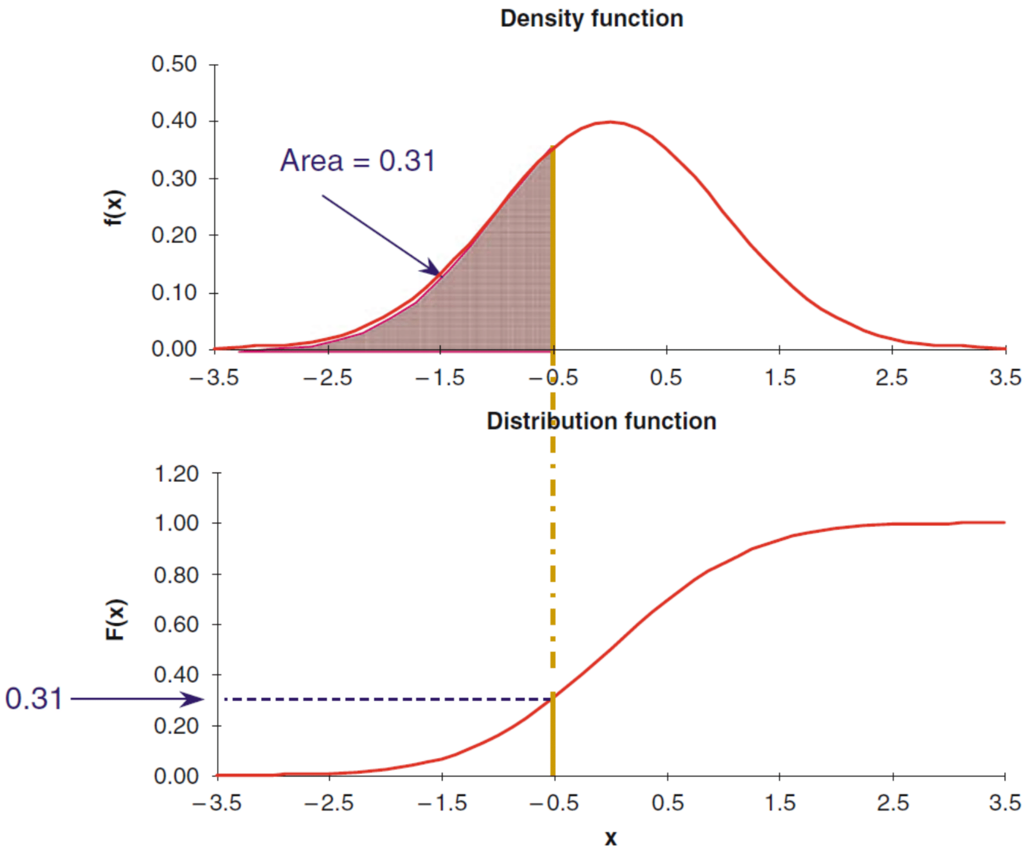

Figure 4.4 The relationship between PDF and CDF

Another way of seeing this relationship is with the following plot, where the value of the area undereneath the density is mapped into a new curve that will represent the CDF. While using this relationship, final sentiment tone of word giving the difference of positive and negative area is calculated using equation (7):

$$S_{\mathrm{T}} = 1 - 2CDF(0) \tag{7}$$

where: $CDF(0) =$ negative area under the density function.

4.3.3. Sentiment Tone Results

Examples of calculated sentiment tone values using two approaches are given on the figures in this section.

4.3.3.1. Positive Terms

The following figures 4.5, 4.6, 4.7, 4.8, include words with positive sentiment tones that have a positive effect on people's minds.

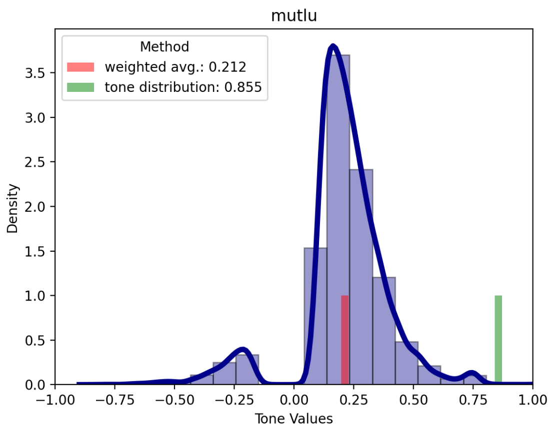

Figure 4.5 Sentiment Tone of ”mutlu”

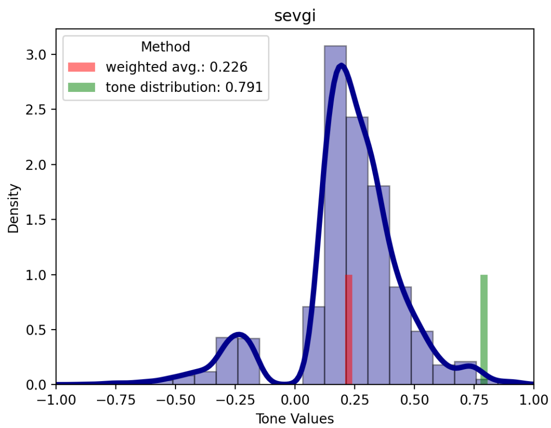

Figure 4.6 Sentiment Tone of ”sevgi”

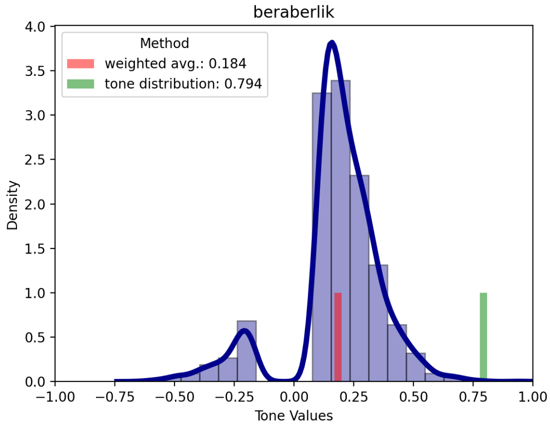

Figure 4.7 Sentiment Tone of ”beraberlik”

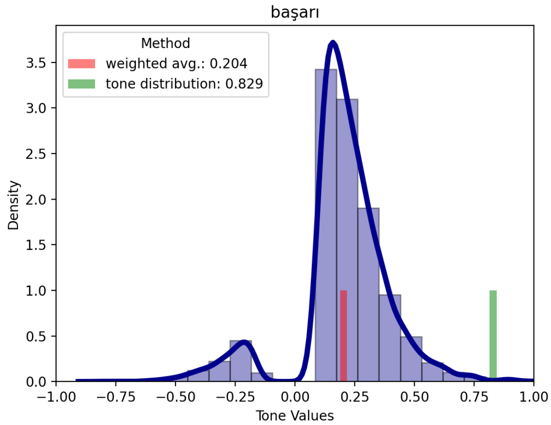

Figure 4.8 Sentiment Tone of ”başarı”

The figures show that the sentiment tones found in both approaches successfully polarized the words as positive. To interpret the tone values found, words with very clear positivity were selected.

The result shows that the sentiment tone found by the weighted average method did not have high positivity, while the tone found by the tone density based method had very strong positivity. This result was obtained because the positive area is quite dominant compared to the negative area as you can see in the figures.

4.3.3.2. Negative Terms

The following figures 4.9, 4.10, 4.11, 4.12, include words with negative sentiment tones that have a negative effect on people’s minds.

The figures show that the sentiment tones found in both approaches successfully polarized the words as negative. To interpret the tone values found, words with very clear negativity were selected. The result shows that the sentiment tone found by the weighted average method did not have high negativity, while the tone found by the tone density based method had very strong negativity. This result was obtained because the negative area is quite dominant compared to the positive area as you can see in the figures.

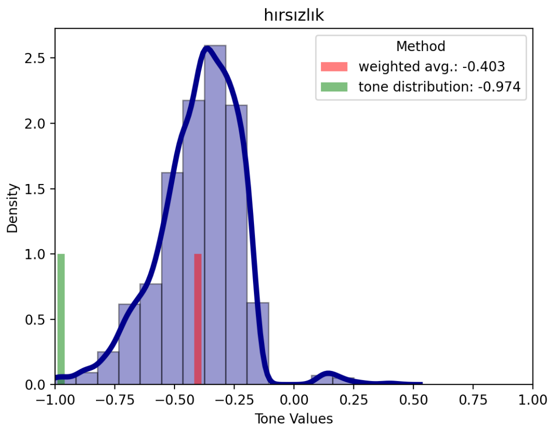

Figure 4.9 Sentiment Tone of ”hırsızlık”

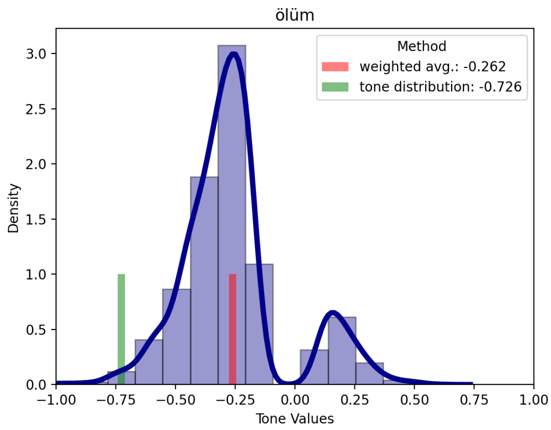

Figure 4.10 Sentiment Tone of ”ölüm”

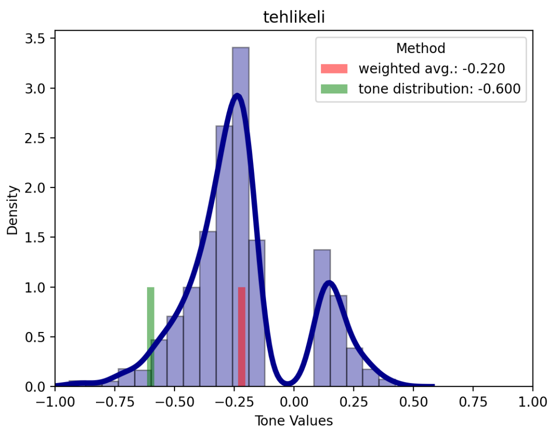

Figure 4.11 Sentiment Tone of ”tehikeli”

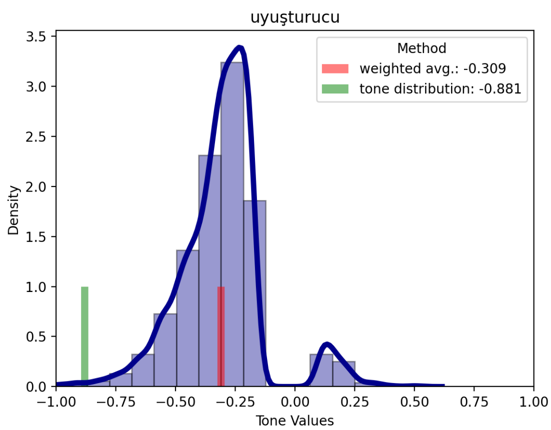

Figure 4.12 Sentiment Tone of ”uyuşturucu”

4.3.3.3. Close to Neutral Terms

The following figures 4.13, 4.14, 4.15, 4.16, 4.17 include words with neutral sentiment tones that have a neutral effect on people’s minds.

The figures show that the sentiment tones found in both approaches are close to neutral. Since the sentiment tone value found by the weighted average method is far from finding value close to the extremes between -1 and 1, tone value is also closer to 0 than the tone density based approach. This result was obtained because the negative area and positive area are almost equal to each other.

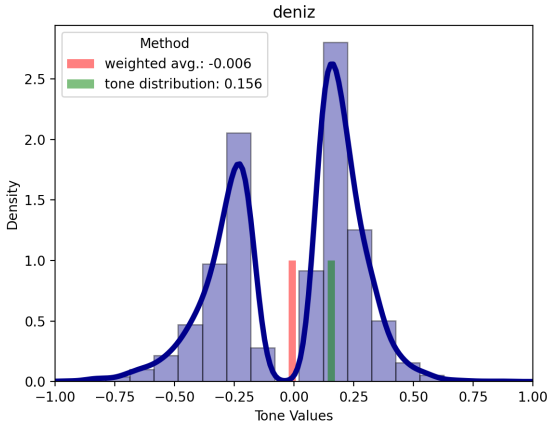

Figure 4.13 Sentiment Tone of ”deniz”

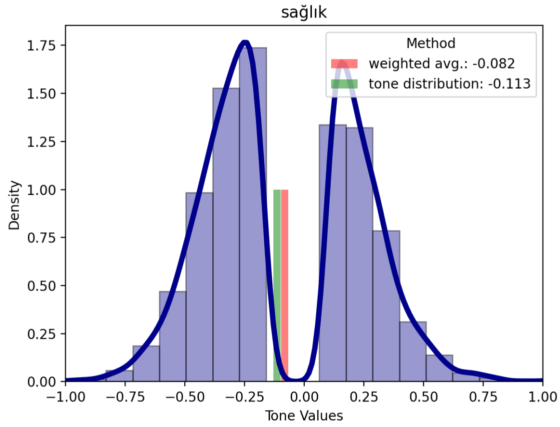

Figure 4.14 Sentiment Tone of ”sağlık”

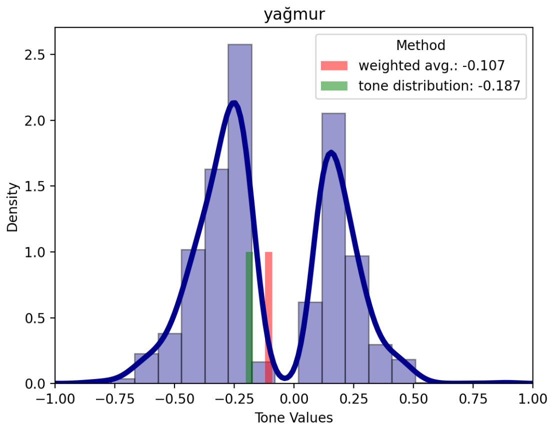

Figure 4.15 Sentiment Tone of ”yağmur”

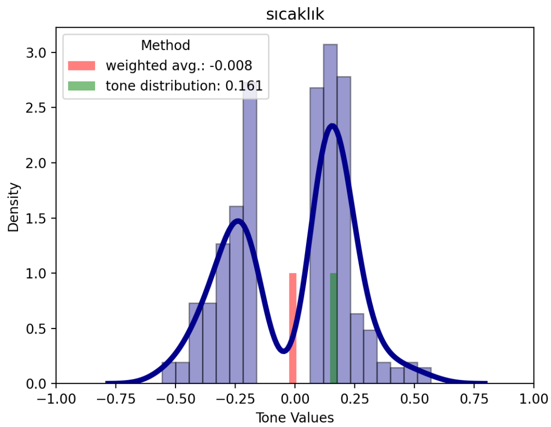

Figure 4.16 Sentiment Tone of ”sıcaklık”

The word ”engelli” has a negative meaning in the human mind. However, the results show that this word has a neutral sentiment tone. The reason for this is that news texts shape word ”engelli” as a positive identity by focusing on the abilities rather than the limitations of this concept. Sample text of this kind of an new was obtained from Ihlas News Agency [43]:

”Engelli vatandaşların çalışma hayatında önünü açmak ve iş fırsatları oluşturmak için çalışmalarını sürdüren Buca Belediyesi Konak Rotary Kulübü işbirliği ile KOSGEB Girişimcilik Eğitimi başlatıyor. 14-17 Ağustos tarihlerinde 4 gün sürecek eğitimler için 9 Ağustos'a kadar Çamlıkule Semt Evi'ne başvurmak gerekiyor. Buca Belediyesi ve Konak Rotary Kulübü işbirliği Buca'da yaşayan ve kendi işini kurmak isteyen engelli vatandaşlar

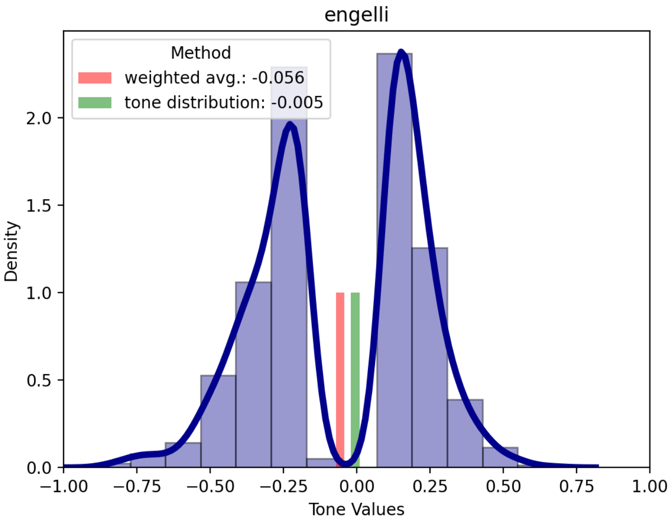

Figure 4.17 Sentiment Tone of ”engelli”

için büyük bir fırsat yarattı. Buca Belediyesi, engelli bireylerin yaşamdan kopmayıp, çalışma hayatına katılması için KOSGEB Girişimcilik Eğitimi düzenleyecek.”

4.4. MLTC (Manually Labeled Turkish Corpus)

For the performance evaluation of the sentiment lexicon, a labeled Turkish corpus is necessary. In order to meet this need, a total of 300 news documents were classified as positive and negative by 3 evaluators. In the evaluation of the evaluators for news, the final polarity of the news was determined on the basis of majority vote. This test corpus was named as MLTC-300.

In Table 4.2, the polarity distributions of MLTC-300 is shown according to the evaluator’s determinations.

TEST SET DISTRIBUTION		
HUMAN_POS	127	42%
HUMAN_NEG	173	58%
TOTAL	300	

Table 4.2 Test Set Distribution

5. EXPERIMENTAL RESULTS

The performance of the lexicons which are developed in two approaches are evaluated and also, comparison of these lexicons and also a general purpose sentiment lexicon SWNETTR++ is developed by Sağlam with a capacity of 49K was analyzed. Results show that the constructed tone density based sentiment lexicon with novel methodology achieves a comparable performance.

The Confusion Matrix was used to see the sentiment lexicon test results. In confusion matrix, manually assigned polarity information for the "Actual" value and the polarity values determined with the help of sentiment lexicons for the "Predicted" value were taken into account.

5.1. Weighted Average Approach

First, the sentiment lexicon created by the weighted average approach was tested. According to the results, the accuracy value of the sentiment lexicon was 87.33%. The values of the confusion matrix of the study are shown in Table 5.1 and the results obtained from the confusion matrix are shown in Table 5.2.

The confusion matrix of sentiment lexicon based on the weighted average approach in Table 5.3 shows that out of 300 messages, 262 were guessed correctly, while 38 were guessed incorrectly. Of 127 positive messages, 94 were successfully predicted, while 33 messages

		Predicted	
		Positive	**Negative**
Actual	**Positive**	94 (TP)	33 (FN)
	Negative	5 (FP)	168 (TN)

Table 5.1 Confusion Matrix for weighted average method

Lexicon with Weighted Avg.	
Precision	0.9495
Recall	0.7402
F1 Score	0.8319
Accuracy	0.8733

Table 5.2 Precision, recall, F1-Score, accuracy values for weighted average method

were incorrectly predicted as belonging to the negative group. Out of 173 negative messages, 168 were predicted to be negative, while only 5 were incorrectly predicted to be positive.

5.2. Tone Density based Approach

The sentiment lexicon created by the tone density based approach was tested. According to the results, the accuracy value of the sentiment lexicon was 88.67%. The values of the confusion matrix of the study are shown in Table 5.3 and the results obtained from the confusion matrix are shown in Table 5.4.

The confusion matrix of tone density based sentiment lexicon in Table 5.3 shows that out of 300 messages, 266 were guessed correctly, while 34 were guessed incorrectly. Of 127 positive messages, 122 were successfully predicted, while 5 messages were incorrectly predicted as belonging to the negative group. Out of 173 negative messages, 144 were predicted to be negative, while only 29 were incorrectly predicted to be positive.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	122 (TP)	5 (FN)
	Negative	29 (FP)	144 (TN)

Table 5.3 Confusion Matrix for Tone Density based Method

Lexicon with Tone Density	
Precision	0.8079
Recall	0.9606
F1 Score	0.8777
Accuracy	0.8867

Table 5.4 Precision, recall, F1-Score, accuracy values for Tone Density based Method

5.3. Merged Method

Combining the results from the two different approaches we implemented, we determined the polarity of the documents as the sign of the sentimental tone, which has a great absolute value. We called it the merged method. According to the results, the accuracy value of the merged method was 89.67%. The values of the confusion matrix of the study are shown in Table 5.5 and the results obtained from the confusion matrix are shown in Table 5.6.

The confusion matrix of merged method in Table 5.5 shows that out of 300 messages, 269 were guessed correctly, while 31 were guessed incorrectly. Of 127 positive messages, 111 were successfully predicted, while 16 messages were incorrectly predicted as belonging to the negative group. Out of 173 negative messages, 158 were predicted to be negative, while only 15 were incorrectly predicted to be positive.

5.4. SWNetTR++

The sentiment lexicon SWNetTR++ was tested. According to the results, the accuracy value of the sentiment lexicon was 89.33%. The values of the confusion matrix of the study are

			Predicted	
			Positive	Negative
Actual	Positive		111 (TP)	16 (FN)
	Negative		15 (FP)	158 (TN)

Table 5.5 Confusion Matrix for Merged Method

Merged Method	
Precision	0.8810
Recall	0.8740
F1 Score	0.8775
Accuracy	0.8967

Table 5.6 Precision, recall, F1-Score, accuracy, values for Merged Method

shown in Table 5.7 and the results obtained from the confusion matrix are shown in Table 5.8.

			Predicted	
			Positive	Negative
Actual	Positive		105 (TP)	22 (FN)
	Negative		10 (FP)	163 (TN)

Table 5.7 Confusion Matrix for SWNetTR++

SWNetTR++	
Precision	0.9130
Recall	0.8268
F1 Score	0.8678
Accuracy	0.8933

Table 5.8 Precision, recall, F1-Score, accuracy, values for SWNetTR++

The confusion matrix of sentiment lexicon SWNetTR++ in Table 5.7 shows that out of 300 messages, 268 were guessed correctly, while 32 were guessed incorrectly. Of 127 positive messages, 105 were successfully predicted, while 22 messages were incorrectly predicted as belonging to the negative group. Out of 173 negative messages, 163 were predicted to be negative, while only 10 were incorrectly predicted to be positive.

5.5. Test Result Evaluation

Table 5.9 shows the results of the confusion matrices together. In the Recall column, which indicates how accurately positive states are predicted in this table, the most successful result was 96.06% for the lexicon based on tone density, while a low percentage was obtained with 74.02% for the weighted average lexicon. In the specificity column, which indicates how accurately the negative states are predicted, the most successful result is 97.11% for the weighted average lexicon, while the lowest result is 83.24% for the tone density based lexicon. Consequently, the weighted average lexicon was more successful in predicting negative news and the tone distribution based lexicon was more successful in predicting positive news.

Since the merged method was tested with the combination of the results of the two methods, the precision and recall values were also between the results of these two approaches, and with the balance of these values, F-score and accuracy were high.

The most successful result is shown for the weighted average lexicon in the Precision column, which is the measure of how many of the positively predicted values are actually positive. The F1 value, which is the harmonic mean of the Precision and Recall results, achieved the highest value in the tone distribution lexicon.

Lexicon	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1 Score
WEIGHTED AVG. BASED	0.8733	0.9495	0.7402	0.9711	0.8319
TONE DENSITY BASED	0.8867	0.8079	0.9606	0.8324	0.8777
MERGED	0.8967	0.8810	0.8740	0.9133	0.8775
SWNetTR++	0.8933	0.9130	0.8268	0.9422	0.8678

Table 5.9 Evaluation Results for 300 News

6. CONCLUSION

As a result of the study, we developed a Turkish sentiment lexicon using a novel methodology: using statistical tone density functions computed using a very large document corpus obtained from mainstream Turkish news agencies. We evaluate the performance of this lexicon by using manually labeled test data and compare it with similar lexicons.While accuracy value of weighted average-based lexicon is 87%, the tone density-based approach is 88% and the combination of this two lexicons give the best accuracy with 89%. Results show that the constructed sentiment lexicon with novel methodology achieves a comparable performance and poses many potential improvement possibilities. This lexicon named as ”slex-tr” and can be accessed from github [44]. We believe, the theories and practices developed in this thesis may also be used in different languages for developing sentiment lexicons in those languages. Henceforth, we expect a wider academical impact in the long run.

As a feature work, we will construct a graph (tree) based structure using the agglutinative morphology of Turkish to store and query the tone values of the words. This will further allow us to study the sentimental weights of different suffixes in the language. We will finally produce a nice and readable visualization of these graphs, which will demonstrate multiple features of the words, such as frequency in corpus, tone value, suffix strength, etc. This visualization will further allow researchers to understand the relations between words and make it possible to extend sentiment studies in the future.

REFERENCES

[1] Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. Plos one, 16(2):, 2021.

[2] Fırat Akba, Alaettin Uçan, Ebru Akcapinar Sezer, and Hayri Sever. Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. In 8th European Conference on Data Miningvolume191, 180–184, 2014.

[3] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining, 2006.

[4] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, 34–35, 2005.

[5] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In Lrecvolume4, 2004.

[6] KR1442 Chowdhary. Natural language processing. Fundamentals of artificial intelligence, 603–649, 2020.

[7] Ronen Feldman. Techniques and applications for sentiment analysis. Commun. ACM, 56(4):, 2013. ISSN 0001-0782. doi:10.1145/2436256.2436274.

[8] http://nebula.wsimg.com/72550c5bec78cf7bc908dbefbd04af8e?AccessKeyId=DFB1BA3CED7E7997D5B1&disposition=0&alloworigin=1, [Online; accessed 14-October-2021].

[9] Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. Decision support systems, 53(4):, 2012.

[10] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177, 2004.

[11] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. Computational linguistics, 37(1):, 2011.

[12] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2):, 2011.

[13] Suheer Al-Hadhrami, Norah Al-Fassam, and Hafida Benhidour. Sentiment analysis of english tweets: A comparative study of supervised and unsupervised approaches. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019.

[14] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In Coling 2010: Posters, 241–249, 2010.

[15] Natalia Loukachevitch and Anatolii Levchik. Creating a general Russian sentiment lexicon. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), , Portorož, Slovenia, 2016.

[16] Seongik Park and Yanggon Kim. Building thesaurus lexicon using dictionary-based approach for sentiment classification. In 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), 39-44, 2016. doi:10.1109/SERA.2016.7516126.

[17] Yi Ding, Bing Li, Yuqi Zhao, and Can Cheng. Scoring tourist attractions based on sentiment lexicon. In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 1990-1993, 2017. doi:10.1109/IAEAC.2017.8054363.

[18] Merve Rumelli, Deniz Akkuş, Özge Kart, and Zerrin Isik. Sentiment analysis in turkish text with machine learning algorithms. In 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), 2019.

[19] Buket Erşahin, Özlem Aktaş, Deniz Kilinc, and Mustafa Erşahin. A hybrid sentiment analysis method for turkish. Turkish Journal of Electrical Engineering & Computer Sciences, 27(3):, 2019.

[20] Bing Liu. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1):, 2012.

[21] Cumali Türkmenoglu and Ahmet Cüneyd Tantug. Sentiment analysis in turkish media. In International Conference on Machine Learning (ICML), 2014.

[22] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. Journal of the American society for information science and technology, 61(12):, 2010.

[23] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3):, 2013.

[24] Yeliz Yengi and Sevinç İlhan Omurca. Distributed recommender systems with sentiment analysis. Avrupa Bilim ve Teknoloji Dergisi, 4(7):, 2016.

[25] http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf, [Online; accessed 2-September-2021].

[26] http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf, [Online; accessed 2-September-2021].

[27] Ahmet Afsin Akın and Mehmet Dündar Akın. Zemberek, an open source nlp framework for turkic languages. Structure, 10(2007):, 2007.

[28] Erik Cambria, Robyn Speer, C. Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining. AAAI Fall Symposium - Technical Report, 14-18, 2010.

[29] Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer. Sentiturknet: a turkish polarity lexicon for sentiment analysis. Language Resources and Evaluation, 50(3):, 2016.

[30] Umut Eroğul. Sentiment analysis in Turkish. Master's thesis, Middle East Technical University, 2009.

[31] Alaettin Ucan, Behzad Naderalvojoud, Ebru Akcapinar Sezer, and Hayri Sever. Sentiwordnet for new language: automatic translation approach. In 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2016.

[32] Fatih Sağlam, Hayri Sever, and Burkay Genç. Developing turkish sentiment lexicon for sentiment analysis using online news media. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), 2016.

[33] Mesut Kaya, Güven Fidan, and Ismail H Toroslu. Sentiment analysis of turkish political news. In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technologyvolume1, 2012.

[34] Fatih Sağlam, Burkay Genç, and Hayri Sever. Extending a sentiment lexicon with synonym–antonym datasets: Swnettr++. TURKISH JOURNAL OF ELECTRICAL ENGINEERING COMPUTER SCIENCES, 27:, 2019. doi:10.3906/elk-1809-120.

[35] Shashank Shekhar Sharma and Gautam Dutta. Sentidraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. Information Processing Management, 58(1):, 2021. ISSN 0306-4573. doi:https://doi.org/10.1016/j.ipm.2020.102412.

[36] Dong Deng, Liping Jing, Jian Yu, Shaolong Sun, and Michael K. Ng. Sentiment lexicon construction with hierarchical supervision topic model. IEEE/ACM

Transactions on Audio, Speech, and Language Processing, 27(4):, 2019. doi:10.1109/TASLP.2019.2892232.

[37] Annet John, Anice John, and Reshma Sheik. Context deployed sentiment analysis using hybrid lexicon. In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019.

[38] Gökhan Yurtalan, Murat Koyuncu, and Çiğdem Turhan. A polarity calculation approach for lexicon-based turkish sentiment analysis. Turkish Journal of Electrical Engineering & Computer Sciences, 27(2):, 2019.

[39] https://github.com/codelucas/newspaper, [Online; accessed 10-October-2021].

[40] https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html, [Online; accessed 15-November-2021].

[41] https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html, [Online; accessed 15-November-2021].

[42] https://bookdown.org/tara_manon/MF_book/continuousrv.html, [Online; accessed 4-September-2021].

[43] https://www.iha.com.tr/izmir-haberleri/bucada-engelliler-icin-ucretsiz-kosgeb.-girisimcilik-egitimi-1767280/, [Online; accessed 24-March-2022].

[44] https://github.com/sulekaraslar/sentiment-lexicon, [Online; accessed 15-September-2022].