Nur Banu ALBAYRAK

M.S. Thesis in Computer Engineering

January - 2011

OPINION AND SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING TECHNIQUES

by

Nur Banu ALBAYRAK

January 2011

OPINION AND SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING TECHNIQUES

by

Nur Banu ALBAYRAK

A thesis submitted to

The Graduate Institute of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

January 2011

Istanbul, Turkey

APPROVAL PAGE

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Assoc. Prof. Veli HAKKOYMAZ
Head of Department

This is to certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist. Prof. Zeynep ORHAN
Supervisor

Examining Committee Members

Assist. Prof. Zeynep ORHAN ______________________

Assist. Prof. Atakan KURT ______________________

Assist. Prof. Özgür ÖZDEMİR ______________________

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Sciences and Engineering.

Assoc. Prof. Nurullah ARSLAN
Director

Date
January 2011

OPINION AND SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Nur Banu ALBAYRAK

M. S. Thesis - Computer Engineering
January 2011

Supervisor: Assist. Prof. Dr. Zeynep ORHAN

ABSTRACT

This study examines the relation between Turkish language usage and psychological states. Writings in Turkish gathered from depressive, non-depressive, anxious or non-anxious people. The writings are analyzed according to usage of following attributes: Each word, groups of words that are mostly used by each diagnoses group, tenses, personal pronouns, verbs and nouns.

Attributes are obtained by using the results of morphological analyses. Morphological analyses of these writings are done through a morphological analysis program called Zemberek (Akın and Akın, 2007). Another program called Weka (Frenk and Witten, 2005) is used to analyze the writings of each person according to the attributes. The analysis process is done via classification methods.

An application is implemented to test the differences in word usage of the diagnosed people. The results of the test application showed that, word usage in Turkish gives many clues about psychological states of people.

Keywords: Natural Language Processing, Opinion Mining, Sentiment Analysis, Morphological Analysis

DOĞAL DİL İŞLEME TEKNİKLERİ KULLANARAK GÖRÜŞ VE DUYGU ANALİZİ

Nur Banu ALBAYRAK

Yüksek Lisans Tezi – Bilgisayar Mühendisliği
Ocak 2011

Tez Yöneticisi Yrd. Doç.Dr. Zeynep ORHAN

ÖZ

Bu çalışma Türkçe'nin kullanımı ile psikolojik durum arasındaki ilişkiyi araştırmaktadır. Depresyonlu, depresyonsuz, anksiyeteli, anksiyetesiz kişilerden Türkçe yazılar toplanmıştır. Yazılar şu özelliklerin kullanımına göre analiz edilmiştir; her bir kelime, her bir tanı grubu tarafından en çok kullanılan kelime grupları, kipler, kişi zamirleri, fiiller ve isimler.

Kullanılan özellikler morfolojik analizler sonucunda elde edilmiştir. Yazıların morfolojik analizleri Zemberek adlı bir morfolojik analiz program kullanılarak yapılmıştır (Akın and Akın, 2007). Weka (Frenk and Witten, 2005) isimli diğer bir program da yazıların bu özelliklere göre analiz edilmesi için kullanılmıştır. Analiz işlemi sınıflandırma metodları ile yapılmıştır.

Tanısı konulmuş kişiler arası kelime kullanımı farklarını test etmek için bir program geliştirilmiştir. Test programının sonuçları, Türkçe'de kelime kullanımının psikolojik durum hakkında bir çok ipucu verdiğini göstermiştir.

Anahtar Kelimeler: Doğal Dil İşleme, Fikir Madenciliği, His Analizi, Morfolojik Analiz

DEDICATION

To my family

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I express sincere appreciation to Assist. Prof. Dr. Zeynep Orhan for her guidance and insight throughout the research.

Thanks go to the Memory Center (Center of Neuropsychiatry) and the academicians that filled the Data Gathering Form for their assistance while gathering the data. Special thanks go to Prof. Dr. Nevzat Tarhan for his advices and interest in the psychological text analysis studies.

I express my thanks and appreciation to my family for their understanding, motivation and patience. Lastly, but in no sense the least, I am thankful to all colleagues and friends who made my stay at the university a memorable and valuable experience.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION	
AUC	Area Under ROC Curve
AW	Anxiety Words
B	Bachelor
BDI	Beck Depression Inventory
BIS/BAS	Behavioral Inhibition System/Behavioral Activation System
BSCT	Beier Sentence Completion Test
CSV	Comma Separated Values
DW	Depression Words
HP	High School or Primary School
IDD-L	Inventory to Diagnose Depression – Lifetime
IE	Information Extraction
IT	Information Technology
LIWC	Linguistic Inquiry and Word Count
MALLET	Machine Learning for Language Learning Toolkit
ML	Machine Learning
MPQA	Multi-Perspective Question Answering
MT	Machine Translation
NAW	Non Anxiety Words
NDW	Non Depression Words

NI	No Information
NLP	Natural Language Processing
OMSA	Opinion Mining and Sentiment Analysis
PANAS	Positive and Negative Affect Scale
PG	Postgraduate
SAI	State Anxiety Inventory
SVM-SVC	Support Vector Machine/Classifier
TAI	Trait Anxiety Inventory
TIPD	Text Investigator for Psychological Disorders

CHAPTER 1

INTRODUCTION

Today, with the improvements in technology, a huge amount of data is accessible for companies, scientists and other people. Especially the widespread use of internet gives many possibilities to mankind. In today's world:

- People do shopping considering forum posts
- Companies measure the customer satisfaction of their products according to product reviews in shopping sites
- People share their everyday life with the people from all over the world by writing blogs

Statistical search results show that there is an upsurge of information sharing especially via internet. A research (Mitchell et al., 2010) in America indicates that, 46% of the Americans say that they get news from four to six media platforms on a typical day; just 7% of them get their news from a single media platform on a typical day. Another research (Fox and Zickuhr, 2009) mentions that in America, 33% of internet users have Twitter or another service account to share updates about themselves or to see updates about others as of September 2009 while, 20% of American internet users have account on status updates services as of December 2008.

The upsurge mentioned above gives rise to many potential scientific applications. Scientists have the chance to access many sources for their studies through internet. They can access any study they need independent from the region. They can share their ideas and even maintain studies together with other scientists anywhere on earth. Finding data for studies also became easy with the improvements in technology, by searching on internet and communicating with other scientists. They can also find prepared data or form their own data by using the information shared on internet. Especially for social sciences many information is available on the web. Interdisciplinary studies between different science branches became easier with the mentioned improvements. This study is one of them.

There are many sources of data to use in scientific studies. Newspapers, televisions, radios, magazines and many other media provide this data. Internet is the major source of this kind of data. Web blogs entries, product reviews, forum posts, update sharing site posts are some of the examples of data that internet provides. News, dialogue records, magazine articles are some of the examples from sources other than internet.

Most of the mentioned kind of data is in text or speech format. Data in these formats are very valuable for social sciences as well as for business world to analyze and get useful results. The results of such analysis are very important for business world in many aspects like better customer satisfaction and more efficient working of employees. With the analysis of the data, companies can also effectively analyze their situations with low expense. Besides, with analysis of data in these formats, many new findings become achievable for different areas of social sciences and for Information Technology (IT) industry.

Psychology is one of these areas that need analysis of these kinds of data in order to obtain new findings and make social observations. With the contributions of the data available via the mentioned media, many observations about the psychological state of large masses can be done. People of any group can be classified by diagnosing their psychological disorders based on old findings, and new findings can be done according to these diagnosed classes. Diagnosing disorders is one of the main objectives in psychology. Diagnosing process can be easier and more efficient with the contributions of text or discourse analysis methods. Analysis of language is needed, for all these purposes.

It is almost impossible to analyze the enormous amount of data manually. Even for fewer data, manual analysis is not efficient. Human analyzers may not be as accurate as automatic analyzers. Manually analyzing texts or discourses is a time consuming task.

Thereby a lot of people and consequently a lot of money are required for manual analysis.

Thus, the importance of computerized methods to analyze that data increases. The hugeness of the amount of the data makes the analysis impossible without computerized methods. Text or discourse analysis is less time consuming with computerized methods. For both academic and business world, computerized methods are efficient and cheaper. Also the results of the computerized analysis are more accurate. James Pennebaker and Cindy Chung mentioned this in their study with these words; “Computerized tools provide efficient and reliable measurement beyond even the most conscientious of human coders.” (Chung and Pennebaker, 2007) Another very time consuming task is to clean the data from redundant information. For all these purposes automated analysis systems are needed to handle the entire data.

1.1. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is the area of computational sciences that deals with language. NLP area explores how to make computers that process, understand, and generate languages. NLP techniques use Machine Learning (ML) algorithms which make computers gain behaviors considering the data they obtain. NLP involves many subareas and tasks. Automatic summarization, discourse analysis, machine translation, relationship extraction, question answering are only a few examples of these areas. The development of these areas results with the development of many other areas in many other fields.

1.1.1. Brief History

NLP studies exist dating back to 1940s. The first application of NLP was a Machine Translation (MT) application developed for breaking enemy codes during World War 2. In 1950 Alan Turing proposed a criterion of intelligence which is now called Turing Test (Turing 1950). The criterion is the ability of a computer to imitate a person in a conversation with a human judge. Hence the evaluation of the intelligence of a computer could be done by using NLP techniques. In 1957 Noam Chomsky introduced an idea (Chomsky 1957) which suggests that there is a universal grammar and all languages have properties of this grammar. In 1966 Automatic Language Processing Advisor Committee (ALPAC) published a report suggesting that ten years long research could not satisfy expectations. After the report, research on NLP reduced too much internationally. The software SHRDLU (Winogard 1971) which is formed in 1970 is one of the succeeding early studies after this recession. It is a software that performs the user's demands in a graphical world of blocks. Roger Schank and some of his colleagues examined human conceptual knowledge. One of them is the study with Robert P. Abelson which is on theory of knowledge systems (Schank and Albelson 1977).

From the late 1980s a revoluation happened with the entering of ML approaches to NLP. The study of Stanley R. Rosenschein and Stuart M. Shieber (Rosenschein and Shieber 1982) is one of the studies that worth mentioning. The study discusses a scheme for syntax-directed translation that mirrors compositional model–theoretic semantics.

With the developments in computer sciences, 1990s became the upsurge period for NLP. Different approaches were examined with contributions of computerized methods. One of the valuable studies in 1990s is the study of Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra (Berger et. al. 1996). The study presents a maximum-likelihood approach for automatically constructing maximum entropy models. They implement the approach efficiently.

A comparative study (Yang and Pedersen, 1997) that examines five different feature selection methods is important with its results. The examination process is done on statistical learning of text categorization. The methods that are examined are:

- Document Frequency
- Information Gain, Mutual Information
- Chi Square Test
- Term Strength

The results of the study showed that Information Gain and Chi Square Test are the best for statistical learning of text categorization.

The study of T. Joachims (Joachims, 1998) which explores the text classifiers learning with the use of Support Vector Machine (SVM) is one of the important studies. The suitability of SVM is proved by analyzing particular properties of learning with text data. It provides both theoretical and empirical evidence while examining the mentioned issue.

Another very important study is the study of Stephen Soderland (Soderland, 1999) that presents a system that is designed to handle different text styles. The system aims to deal with the separate set of rules requirement problem of Information Extraction (IE) systems, by learning text extraction rules automatically. The system aims to handle different text styles ranging from highly structured to free.

The survey of Fabrizio Sebastiani (Sebastiani, 2002) provides a broad scanning of text categorization approaches within ML. The study focuses on three problems: Document Presentation, Classifier Construction and Classifier Evaluation.

1.1.2. State of the Art

Today NLP is a broad area with branches related to many areas. Many studies are being maintained today with the contributions of NLP techniques. Some state of the art examples are as follows.

One of the recent studies (Liu et al., 2007) of NLP addresses the problem of detecting low quality product reviews in opinion summarization. The study explores namely the informativeness, readability, and subjectiveness of the product reviews by using a classification method. The approach is implemented to opinion mining task and a two-stage framework is formed. The study explores 4909 reviews that are labeled by two annotators. The average accuracy of the approach is 78%.

Another study (Jin et al., 2007) which uses classification method, examines sensitive classification of web pages for content advertising. The study may be very valuable for the advertisement sector.

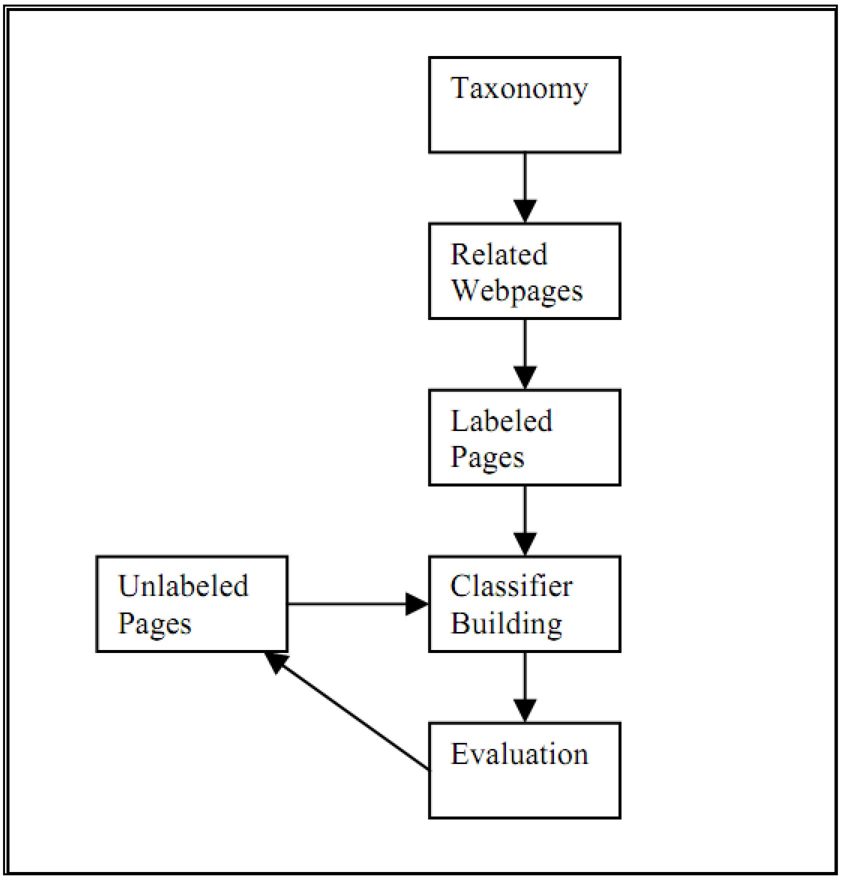

Figure 1.1 The Methodology of Sensitive Webpage Classification for Content Advertising (Jin et al., 2007)

Figure 1.1 shows the methodology of the study. About 200 web pages are labeled according to some categories. These pages are randomly divided as training and validation sets. Some other pages are collected and used as unlabeled and labeled pages to form the evaluation set. The average accuracy is for hierarchical classifiers is 57% and for binary classifiers 78.5%.

A study (Hu et al., 2007) that worths mentioning does blog summarization considering the representative sentences of the blog post comments. Figure 1.2 shows the methodology of the study. Features of the application are appearance of a word in at least one of the comments, comment frequency, and term frequency. The application computes a representativeness score for each sentence by using the features, and then selects the sentences with higher scores than a threshold. The average precision value for different evaluation measures that are used in the study is 48.5%.

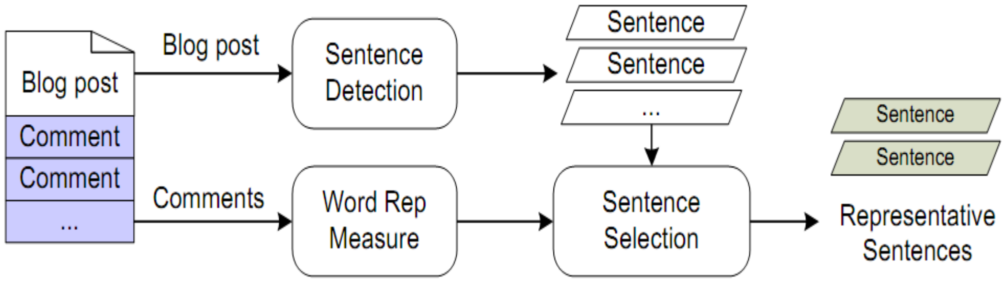

Figure 1.2 Methodology of Comments Oriented Blog Summarization by Sentence Extraction (Hu et al., 2007)

Another study (Ghose and Ipeirotis, 2007) that is very important for evaluation of reviews presents a system for ranking reviews predicting their usefulness and impact. The system uses classification method for the ranking process. The ranking is done considering two parameters. These are ranking considering consumers and ranking considering manufacturers. Consumer oriented ranking examines expected helpfulness and manufacturer oriented ranking examines expected affects on sales. N-gram attributes are used in the classification process. Application first predicts the subjectivity, and then, estimates the relation between sales rank and subjectivity in review. The relation between votes and subjectivity of the review is estimated too. These attributes are used to measure the helpfulness and impact of the review. 1000 reviews are classified by human coders. The reviews are classified as informative or not and impact or not. The F-measure of estimating the usefulness of a review is 0.85.

The study of A. Abbasi and H. Chen on affect intensity analysis of dark web forums (Abbasi and Chen, 2007), introduces a system for analyzing extremist groups' forum postings. The attributes of the analysis are terms that represent feelings like happiness, anger, sadness, horror, etc. Figure 1.3 shows the structure of the system. The system consists of two parts. The first part is the creating of an affect lexicon. The second part is execution of an intensity analysis technique that can be used to perform affect analysis. Messages from 16 forums are collected and analyzed by linear regression analysis.

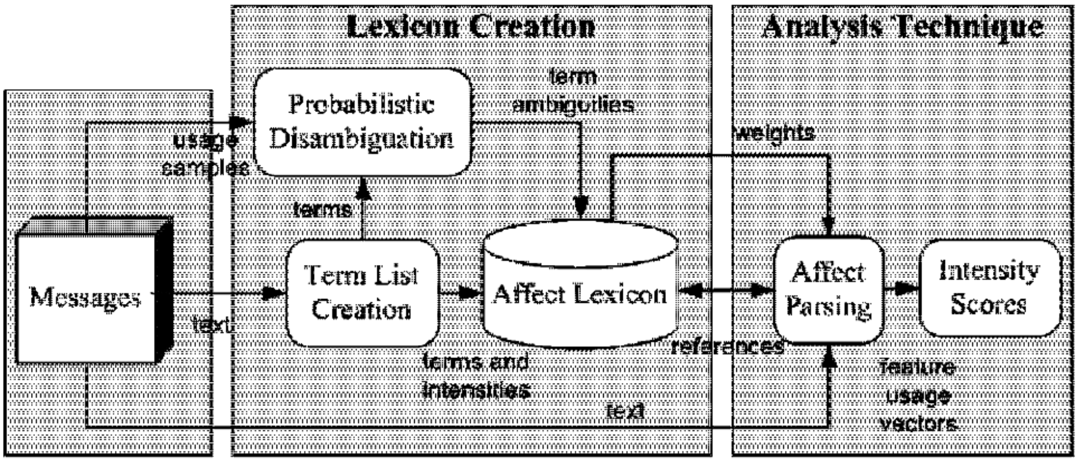

Figure 1.3 Affect Analysis System Design (Abbsai and Chen, 2007)

1.2. OPINION MINING AND SENTIMENT ANALYSIS

Opinion Mining and Sentiment Analysis (OMSA) term that consists of two areas deals with opinions, sentiments and subjectivity:

- Opinion Mining
- Sentiment Analysis

OMSA systems basically classify texts or discourses considering specific attributes. Classification process may have various categories. One style can be:

- Positive
- Negative
- Neutral

Another one can be:

- Subjective
- Objective

Alternatively, the classes may be formed according to other attributes.

Referring to a recent survey (Pang and Lee, 2008) on OMSA, early projects on beliefs can be counted as forerunners of the area. Two examples of these studies are mentioned in the survey:

- Subjective Understanding: Computer Models of Belief Systems (Carbonell, 1979).
- Beliefs, Points of View, and Multiple Environments (Wilks and Bien, 1983).

Today many OMSA studies have been explored. Use and development of OMSA techniques has an upsurge with rise of ML and NLP. Some example studies are given as follows.

A recent study (Ding et al., 2008) of OMSA presents a holistic lexicon-based approach to opinion mining. It examines determination of semantic orientations (positive, negative, neutral) of opinions. The data used for the study consists of reviews of products. Opinion words around each product feature in a review sentence are used as attributes. The approach scores words as negatives and positives, checking negation words and but clauses; and handling context dependent opinions with conjunction rules. Part of speech tags are used in order to determine possible different usages of each word. A system called NLProcessor (Infogistics, 2000) used for the part of speech tagging process. 445 reviews of 8 different products are collected. A holistic lexicon-based approach is used by using external evidences and linguistic structures. The average precision value of results with different attributes is approximately 91%.

A study (Pang and Lee, 2008) on spam analysis uses product reviews as data and uses novel attributes that have different aspects. The attributes used for the analysis are grouped as review centric features, reviewer centric features and product centric features. Some examples to review centric features are percentage of positive and negative opinion bearing words, and cosine similarity of the review and product features. An example to reviewer centric features is average rating given by the reviewer. Product centric feature examples are price and sales rank. Area Under ROC Curve (AUC) criteria is selected to evaluate the classification results. The average AUC value for different attributes is approximately 98%.

Another important study (Breck et al., 2007) on OMSA presents an approach for identifying expressions of opinion. The approach presented in the study is for identifying opinion expressions that uses conditional random fields. A corpus called Multi-Perspective Question Answering (MPQA) (Wiebe et al, 2005) is used. The corpus consists of 535 newswire documents. They used 10-fold cross validation for the classification process. Some utilities are used in order to accomplish the tasks like part of speech tagging or parsing. A linear chain conditional random field model is used for learning method. For this purpose Machine Learning for Language Toolkit (MALLET) (McCallum 2002) is used. The results are reported according to two predicates. These are called exact and overlap.

- **exact(c,p)** is true when c and p are the same spans,
- **overlap(c,p)** is true when c and p spans overlap where c and p shows the correct and predicted expression spans, respectively.

One of the recent studies (Fukuhara et al., 2007) examines the sentiment analysis of social events. The study aims analyzing trends of people's sentiments along with timeline. The data of the study consists of texts with timestamp, such as web blogs and news articles. The attributes of the study is sentiment phrases such as "happy", "good", "bad". Eight categories of sentiment phrases are detected. These are:

- Anxiety
- Sorrow
- Anger
- Happiness
- Suffering
- Fatigue
- Complaint
- Shock

Texts with time stamp are gathered for each day. For a chosen sentiment category following steps are executed to make a topic graph. Keywords are extracted by using a utility. For each keyword an average correlation is calculated between the keyword and sentiment phrases in the sentiment category. Top n keywords are extracted according to some specific criteria. The keywords are put into clusters and a graph is generated. The sentiment graph is generated by considering frequency values of the sentiment phrases. (See Figure 1.4 for an overview.)

The system called Crystal (Kim and Hovy, 2007), analyzes predictive opinions on the web. The study represents a system that examines the prediction of election results from web user opinions on election prediction sites. N-gram feature patterns are used with Support Vector Machines (SVM) for the analysis. The system approach consists of three steps:

- Feature generalization
- Classification using SVM
- SVM result integration

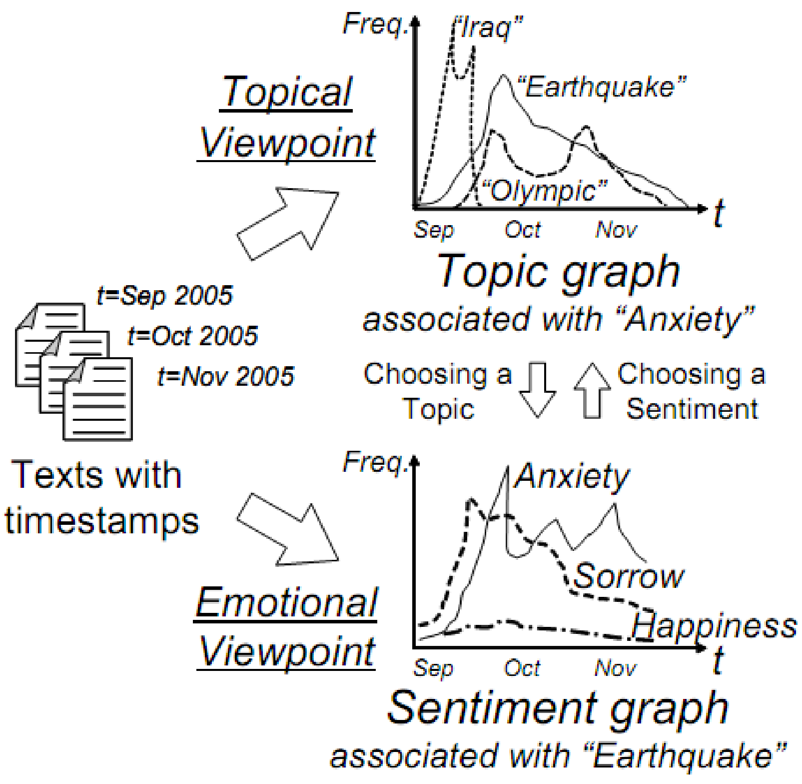

Figure 1.4 Overview of Temporal Sentiment Analysis (Fukuhara et al., 2007)

Two corpora of old election predictions are used. These corpora have 4858 and 4680 messages, respectively. 10-fold cross validation method is used for classification. The accuracies for different n values of n-grams are above 65% (See Table 1.1 for the accuracy of the analysis of the Crystal system).

There are many fields that may move forward with the contributions of OMSA approaches and applications as shown in the examples. One of these fields is psychology. Psychology deals with tests, texts and discourse while diagnosing patients. OMSA approaches may help analyzing these data while maintaining the diagnosis process.

Table 1.1 Accuracy Values of the Analysis of the Crystal System

Features	**Accuracy (%)**
uni	72.03
bi	71.81
tri	69.57
four	67.64
uni + bi	72.93
uni + tri	72.20
uni +four	72.84
bi + tri	72.26
bi + four	72.17
uni + bi + tri	73.07
uni + bi + four	72.30
uni + tri +four	72.30
bi + tri +four	72.62
uni + bi + tri + four	73.01

1.3. RELATED WORK ON PSYCHOLOGY

Starting from the early 1900s natural language is accepted as a potentially diagnostic of a person’s psychological state. Before the widespread use of computers, psychologists put forward the idea that language use is potentially a strong indicator of psychological state of a person.

Freud is one of the forerunners of the idea with his study in 1901 (Freud, 1901). He associates the slips of tongue with people's deeper motives. McClelland's ideas are also parallel to other scientists who argued the relation between natural language and psychology. He has many studies proving the relation. One of these is in 1948 with Atkinson (Atkinson and McClelland, 1948). They analyze written stories of people who had food deprivation for 1, 4 or 6 hours. Gottschalk and Gleser contribute the idea with their study in 1969 (Gottschalk and Gleser, 1969). They suggest that valuable information of one's psychological state can be obtained by analyzing written text or discourse.

Psychological text analysis rapidly develops following these initial studies. The use of computerized methods has a great contribution to the growth of the quantity of studies on this area. Today researchers and scholars of psychology and psychiatry have the opportunity to handle huge amounts of data through computerized methods. It is possible to analyze data of large mass of people with the contribution of computers. In one of his articles Pennebaker (Pennebaker, 2007) mentions that researchers of health psychology are now standing at the threshold of a new era that suggests some novel ways to think about words, natural language and narrative. Pennebaker and many researchers develop the area by employing the opportunities of computers.

In the following paragraphs some of the recent studies are examined. These studies and their results prove the contributions of text analysis to psychology.

Rude, Gortner, and Pennebaker (Rude et al., 2004) analyze texts written by students and confirmed previous findings about linguistic symptoms of depression. In their study, a group of college students that consists of currently-depressed, formerly-depressed and never-depressed students were asked to write essays. These essays were examined in order to detect language differences between these groups and state the cognitive operations associated with depression and depression vulnerability. A computer program is generated to count the incidence of the words in predesigned categories.

Sample group of the study consists of undergraduates from introductory psychology classes at the University of Texas at Austin. There are 31 (29 female) currently depressed, 26 (20 female) formerly depressed, and 67 (47 female) never depressed participants.

The Beck Depression Inventory (BDI) (Beck et al., 1961) and the Inventory to Diagnose Depression-Lifetime (IDD-L) (Zimmerman and Coryell, 1987) are used to measure depression levels of the groups. These inventories are widely used self-report inventories which measure current depression level and previous depression episodes. BDI is used to measure the depression levels. IDD-L is used to determine prior episodes of depression.

The essays are analyzed by Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) program. LIWC is a program that analyses texts according to some categories. The linguistic dimensions chosen for the study are:

- first person singular (I, me, my)
- first person plural (we, us, our)
- social references (mention of friends, family etc.)
- negatively valenced (gloom, fight etc.) or positively valenced (joyful, accept etc.) words

BDI scores confirmed that currently-depressed people score higher than formerly- and never-depressed ones. IDD-l scores also confirmed the significant difference between never- and formerly-depressed people. Unexpectedly, the formerly-depressed group reported greater past depressive symptoms than the currently-depressed group. (See Table 1.2 for detailed information.) In Table 1.2 linguistic dimensions are shown as mean percentage of words in a given linguistic category out of the total number of words used in the essay.

Table 1.2 Means and (Standard Deviations) on Depression Measures and Linguistic Dimensions for the Three Diagnostic Groups

	Currently Depressed (n=31) **Mean (SD)**	**Formerly Depressed (n=26)** **Mean (SD)**	**Never Depressed (n=67)** **Mean (SD)**
Age	17.97 (0.41)	18.96 (2.82)	18.78 (4.03)
BDI	20.05 (6.39)	4.17 (1.36)	3.58 (1.37)
IDD-L	25.60 (17.21)	40.59 (13.14)	1.18 (2.21)
Linguistic Dimensions:			
First Person Singular (I, me, my)	12.17 (2.91)	10.76 (3.10)	10.76 (2.51)
"I" Only	8.50 (2.22)	7.88 (2.15)	7.27 (2.02)
Negatively Valenced Words	2.92 (1.26)	1.70 (0.80)	1.63 (0.91)
Positively Valenced Words	2.64 (1.28)	3.49 (1.30)	3.12 (1.45)
Social Words	6.32 (2.17)	5.89 (2.53)	5.78 (2.83)

One of the most important studies is LIWC (Pennebaker et. al., 2007). LIWC is a text analysis program that is for studying the emotional, cognitive and structural components presented in verbal and written speech of people. LIWC uses a dictionary that consists of words of categories. (See Table 1.3 for a sample from LIWC dictionary.) The dictionary includes almost 4500 words and word stems. The LIWC program executes the word count process and gives the word count results according to the categories, considering the usage of words and word stems. The study is used as utility for many interdisciplinary studies on psychology and computational linguistics.

Table 1.3 Some Sample Categories and Some Information from LIWC Dictionary

Category	Abbrev	Examples	Words in Category
Total Function Words	funct		464
Total Pronouns	pronoun	I, them, itself	116
Personal Pronouns	ppron	I, them, her	70
Social Processes	social	mate, talk, they, child	455
Family	family	daughter, husband, aunt	64
Friends	friend	buddy, friend, neighbor	37
Humans	human	adult, baby, boy	61

Alex S. Cohen and some other professionals (Cohen et al., 2008) prove the possibility of measuring personality using lexical analysis of natural speech. They use computerized methods to execute lexical analysis. Subjects of the study consist of 35 men and 33 women from a large public university. Speech of each individual about any topic is recorded for 3 minutes. Then subjects completed some personality instruments. The speech records are analyzed by computerized methods and evaluated considering the personality instruments.

They measure two models of personality. One of them is Positive and Negative Affect Scale (PANAS) (Watson and Clark, 1999) which measures separate positive and negative affectivity traits. The other one is Behavioral Inhibition System/Behavioral Activation System (BIS/BAS) (Carver and White, 1994) scale which assesses individuals' enduring motivations relating with their environment. They analyze the transcripts of the speech records by using LIWC program.

1.4. MOTIVATION

This study aims to:

- Contribute to the NLP studies for Turkish language
- Be an interdisciplinary study of computational linguistics and psychology
- Contribute to psychological researches of Turkish language usage

Because:

- There are not enough studies for Turkish on NLP
- NLP studies may enable many new findings for psychology as well as other social sciences

For these purposes:

- Data are gathered from healthy and depression or anxiety diagnosed people
- A system that enables psychological analysis of writings is implemented

CHAPTER 2

EXPERIMENTAL SETUP

This study deals with language processing. Therefore data as text or speech is required. Textual data format is chosen to eliminate the time consuming speech to text conversion. Two types of data are gathered:

- writings plus inventory results
- writings only

These two parts are required since many difficulties are experienced while gathering the data. It is hard to access the data of diagnosed people from the psychologists and psychology departments. Therefore this process started by gathering writings and two very common inventory answers BDI (Beck et al., 1961) and STAI (Spielberg et. al., 1970) from the chosen instances. The scores of the inventories are calculated considering the answers of the instances. Instances are then classified according to their scores. Some professionals agreed to help us finding more data, while data gathering process was continuing. They provided writings and sentence completion test answers of already diagnosed people. So, the second part of the data is formed by the valuable contributions of these professionals.

2.1 METHODOLOGY

This part gives detailed information about the methodology. This study started with data gathering process, as shown in Figure 2.1. Detailed information on data and gathering process is given in section 2.3. Data gathering process continued simultaneously with the development of the psychological disorder investigation software. The program is called Text Investigator for Psychological Disorders (TIPD).

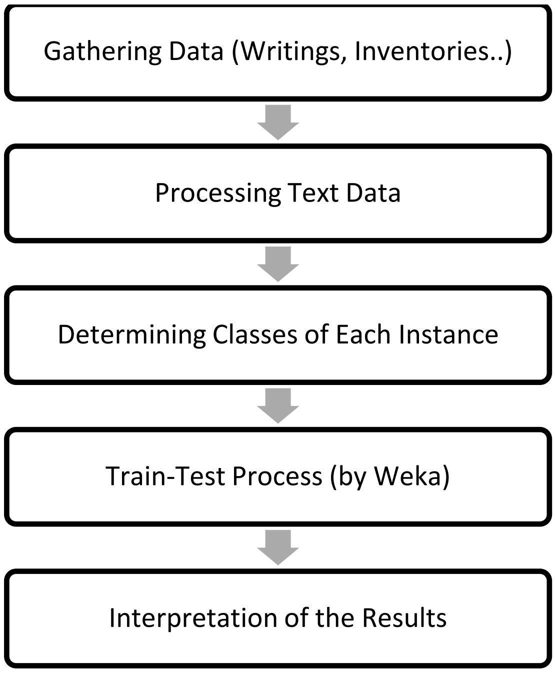

Figure 2.1 Processes

The program provides the opportunity to carry out the procedures of

- **Obtaining Raw Words:** Extracting raw version of each word from whole text.
- **Analyzing Words:** Analyzing each word by using classes of Zemberek.
- **Obtaining Word Usage Results:** Finding word usage results of each instance by using word count techniques.
- **Determining Classes:** Determining the class of each instance according to the inventory results.
- **Train and Test Process:** Training and testing Weka classifiers with the data that consists of words and affixes.
- **Presenting results:** Presenting the results of the testing process.

The writings and inventory answers of the instances are derived in manuscript. These manuscripts are manually translated into digital texts. The multiple writings of each instance merged in one plain text. Additionally he inventory answers of the instances in first part are stored in txt files. The data that are stored in txt format are directly processed by TIPD program. TIPD program divided the writings into sentences and then words. All these words are analyzed by using Zemberek classes (in Java language). Root and affixes of each word are obtained as a result of the analysis. These roots and affixes are used for calculating word and affix frequency. The frequencies are given as input to the classification process that is carries out by using Weka. The results of the classification process are explored in Chapter 3.

2.2 INSTANCES

There are totally 88 instances that are used for this study. Among these, 55 of them belong to the academicians and from the first part of the data. The rest of them, namely 33 instances, form the second part of the data. The instances are chosen from people that are diagnosed with depression and/or anxiety, since this study tries to detect the word usage that distinguishes the depression and/or anxiety diagnosed people.

Table 2.1 shows the diagnoses comparison of depression and anxiety. One of the dimensions of the table shows diagnoses of depression and the other shows diagnoses of anxiety. The instances are evenly distributed as depressed and non depressed, 44 and 44 respectively. Similarly the numbers of instances who are diagnosed and not diagnosed with anxiety are 47 and 41, respectively. There are 10 instances that are not diagnosed with depression or anxiety. 13 of the instances are diagnosed with both depression and anxiety. 34 instances are diagnosed with only anxiety. There are 31 instances that are diagnosed with only depression.

Table 2.1 Comparison of Anxiety and/or Depression Diagnosed Instances

Diagnosis	Anxious	Non-Anxious	Total
Depressive	13	31	44
Non-Depressive	34	10	44
Total	47	41	88

Table 2.2 shows the distribution of depression and anxiety over genders. The table shows the number of instances from each diagnosis and each gender. There are 45 male and 43 female instances. There are 24 depression diagnosed and 23 anxiety diagnosed instances among female instances. 19 female instances have no depression and 20 female instances have no anxiety. There are 20 depression diagnosed and 24 anxiety diagnosed instances among males. 25 of the male instances have no depression and 21 of them have no anxiety.

Table 2.2 Distribution of Diagnoses over Genders

Gender	**Depressive**	**Non-Depressive**	**Anxious**	**Non-Anxious**
Female	24	19	23	20
Male	20	25	24	21
Total	44	44	47	41

Table 2.3 shows the distribution of diagnoses over educational status. The table shows the number of instances from each educational status and each diagnoses group. Educational status of the instances varies as follows: 56 of the instances are postgraduate (PG), 12 of them are bachelor (B), 17 of the instances are high school or primary school (HP) and there are 2 instances with no information (NI) about their educational status.

Table 2.3 Distribution of Educational Status over Diagnoses

Diagnosis	**PG**	**B**	**HP**	**NI**	**Total**
Depressive	13	11	18	2	44
Non-Depressive	44	0	0	0	44
Anxious	40	2	5	0	47
Non-Anxious	17	9	13	2	41

Table 2.4 shows the distribution of educational status over gender. Table shows the number of instances from each educational status and each gender. There are 24 PG, 4

B, 13 HP, and 2 NI in female instances. There are 33 P, 7 B, and 5 HP among male instances.

Table 2.4 Distribution of Educational Status over Gender

Gender	PG	B	HP	NI	Total
Female	24	4	13	2	43
Male	33	7	5	0	55
Total	57	11	18	2	88

2.3 EXPERIMENTAL DATA

2.3.1. First Part of the Data

The inventories that are used in the first part are BDI (Beck et. al., 1961) and STAI (Spielberg et. al., 1970). The instances were asked to write their thoughts on four topics along with the inventories. A form, which consists of these four topics and the inventories, is given to each instance. A sample form is provided in APPENDIX A. Textual data is acquired from the essays written by these instances and the information used for classification of them is calculated from their inventory scores.

In the first part, four topics are given to the instances and they wrote their thoughts about them. These topics are:

Topic 1: Thoughts about the academic life.
Topic 2: The most terrifying event experienced
Topic 3: Thoughts about the current status of the world, family and self.
Topic 4: Comparison of his/her past and present.

These four topics are chosen consulting the psychologists who assisted to this study. In this manner, the topics could be determined in a way to be explored from different aspects. Primary material of this study is the language, specifically the words. All of the analysis and classification process is done according to word usage. Writings are analyzed considering the usage of different kinds of words (nouns, verbs, etc.). Hence it is important for each of the writings to involve different kinds of words.

These four topics are chosen in such a manner that the writings have the potential to be written with different kinds of words. For example it is expected that academicians would like to express their selves more with Topic 1. Although Topic 1 is expected to involve both negative and positive words, the Topic 2 is expected to have more negative words. On the other hand Topic 3 and Topic 4 provide the use of different pronouns and tenses.

2.3.2. Second Part of the Data

The second part of the data consists of writings and sentence completions of people who are diagnosed by psychologists with depression and/or anxiety. These data are gathered by the assistance of psychologists and psychiatrists. Some of the data consist of answers to Beier Sentence Completion Test (BSCT) and the remaining consists of answers to two questions. Only the completion parts of the BSCT are considered for the data.

These questions are:

- Q1: What do you feel about yourself?
- Q2: How do you spend your day?

These two essays are considered and evaluated as one essay due to aforementioned reasons. BSCT is a form that consists of incomplete sentences to complete. Patients complete these sentences according to their feelings and thoughts, and then professionals evaluate them. In this study the portions that are completed by the instances are used and considered and evaluated as an essay. (See Table 2.5 for some example items of BSCT.)

Table 2.5 Some Sample Incomplete Items of BSCT and English Versions

1 Gelecek bana …	1 The future is…
2 Çocukken …	2 When I was a child…
3 Annelerin iyisi …	3 The best of mothers is…
4 İşim …	4 My work is…
5 En büyük kabahat …	5 The worse fault is…

Merging the writings in more than one part may not cause any problem because eventually the handled data consists of words. Since the data is analyzed by word and affix counts the final format of the merged sentence completions is appropriate to the study.

All of the essays and sentence completion test answers of each instance are considered and evaluated as one essay. This consideration is possible because the attributes used in classification process are words. Through this consideration all of the mentioned specifications are collected together, in one file for each instance.

2.3.3. Word Usage Values

Table 2.6 shows the average and standard deviation values of word usage (with recurrence) in different diagnoses groups. Average number of words for the writings of the instances that are contributed to this study is 223.29. Standard deviation of the word usage average for all instances is 147.32. Instances that are diagnosed with depression used words with an average of 197.97 while instances that are not diagnosed with depression used words with an average of 248.61. Standard deviations of these average values are 133.72 and 157.23, respectively. Word usage of instances that are diagnosed with anxiety has an average of 164.83 while word usage of instances that are not diagnosed with anxiety has an average of 176.19. Standard deviations of these two groups' average values are 177.97 and 80.78, respectively. The instance groups of depression and anxiety diagnosed, depression diagnosed but not anxiety diagnosed, not depression diagnosed but anxiety diagnosed and not depression nor anxiety diagnosed have word usage values with the averages of 241.92, 179.54, 272.97 and 165.80, respectively. Standard deviations of these average values are 201.63, 90.43, 170.56 and 39.80, respectively.

Table 2.6 Average (AV) and Standard Deviation (SD) Values of Word Usage with Recurrence in Different Instance Groups

Diagnosis	AV	SD
Depressive	197.97	133.72
Anxious	164.83	177.97
Non-Depressive	248.61	157.23
Non-Anxious	176.19	80.78
Depressive and Anxious	241.92	201.63
Depressive and Non-Anxious	179.54	90.43
Anxious and Non-Depressive	272.97	170.56
Non-Anxious and Non-Depressive	165.80	39.80
All of the Instances	223.29	147.32

Table 2.7 shows the average and standard deviation values of word usage (without recurrence) in different diagnoses groups. There are totally 2190 unique words when all texts are considered. The average of unique words used by all of the instances is 121.59 with a standard deviation value of 57.07. Instances that are diagnosed with depression used unique words with an average of 118.25 while instances that are not diagnosed with depression used unique words with an average of 124.09. Standard deviations of these average values are respectively 67.42 and 52.49. Word usage of instances that are diagnosed with anxiety has an average of 116.45 while word usage of instances that are not diagnosed with anxiety has an average of 126.04. Standard deviations of these two groups' average values are respectively 44.63 and 69.02. The instance groups of depression and anxiety diagnosed, depression diagnosed but not anxiety diagnosed, not depression diagnosed but anxiety diagnosed and not depression nor anxiety diagnosed have word usage values with the averages of 137.76, 106.09, 134.67 and 104.10 respectively. Standard deviations of these average values are respectively 89.47, 42.61, 57.28 and 23.29.

Table 2.7 Average (AV) and Standard Deviation (SD) Values of Word Usage without Recurrence in Different Instance Groups

Diagnosed with	AV	SD
Depressive	118.25	67.42
Anxious	116.45	44.63
Non-Depressive	124.09	52.49
Non-Anxious	126.04	69.02
Depressive and Anxious	137.76	89.47
Depressive and Non-Anxious	106.09	42.61
Anxious and Non-Depressive	134.67	57.28
Non-Anxious and Non-Depressive	104.10	23.29
All of the Instances	121.59	57.07

2.4 INVENTORIES

BDI is one of the main assessment tools for depression which is used by health care professionals and researchers in psychology. BDI (Beck et. al., 1961) is known with its success and stability in obtaining more accurate results and measurements of intensity, severity, and depth of depression. It involves 21 items and these items are rated on a 4 point scale. Points range from 0 to 3. The maximum score is 63. There is a revised version of BDI which is called BDI 2 (Beck et al., 1996), but it has more detailed items to measure the severity of depression for patients that need hospital care. We do not need such details therefore we use BDI. (See Table 2.8 for some sample items of BDI.)

There are four levels of depression that BDI identifies according to the total score:

- Minimum
- Mild
- Moderate
- Severe

Table 2.8 Some Example Items of BDI and the English Versions

1. a. Kendimi üzüntülü ve sıkıntılı hissetmiyorum. b. Kendimi üzüntülü ve sıkıntılı hissediyorum. c. Hep üzüntülü ve sıkıntılıyım. Bundan kurtulamıyorum. d. O kadar üzüntülü ve sıkıntılıyım ki artık dayanamıyorum.	1. a. I do not feel sad b. I feel sad c. I am sad all of the time and I cannot snap out of it d. I am so sad or unhappy that I cannot stand it
2. a. Gelecek hakkında mutsuz ve karamsar değilim. b. Gelecek hakkında karamsarım. c. Gelecekten beklediğim hiçbir şey yok. d. Geleceğim hakkında umutsuzum ve sanki hiçbir şey düzelmeyecekmiş gibi geliyor.	2. a. I am not particularly discouraged about my future b. I feel discouraged about my future c. I feel I have nothing to look forward to d. I feel the future is hopeless and that things cannot improve
3. a. Kendimi başarısız bir insan olarak görmüyorum. b. Çevremdeki birçok kişiden daha çok başarısızlıklarım olmuş gibi hissediyorum. c. Geçmişe baktığımda başarısızlıklarla dolu olduğunu görüyorum. d. Kendimi tümüyle başarısız biri olarak görüyorum.	3. a. I do not feel like a failure b. I have failed more than the average person c. As I look back on my life, all I can see is a lot of failures d. I feel I am a complete failure as a person
4. a. Birçok şeyden eskisi kadar zevk alıyorum. b. Eskiden olduğu gibi her şeyden hoşlanmıyorum. c. Artık hiçbir şey bana tam anlamıyla zevk vermiyor. d. Her şeyden sıkılıyorum.	4. a. I get as much satisfaction out of things as I used to b. I don't enjoy things as I used to c. I don't get real satisfaction out of anything anymore d. I am dissatisfied or bored with everything
5. a. Kendimi herhangi bir şekilde suçlu hissetmiyorum. b. Kendimi zaman zaman suçlu hissediyorum. c. Çoğu zaman kendimi suçlu hissediyorum. d. Kendimi her zaman suçlu hissediyorum.	5. a. I don't feel particularly guilty b. I feel guilty a good part of the time c. I feel quite guilty most of the time d. I feel guilty all of the time

Score range of minimum class is 0 to 13, mild class is 14 to 19, moderate class is 20 to 28 and severe class is 29 to 63. (See Table 2.9 for the depression levels of BDI according to the score ranges.) In this study, instances with scores higher than 14 are considered as depressed. Since the classification process is done in binary format binary classes are required. As a consequence the four levels of depression levels are reduced to two by composing mild, moderate and severe levels. Binary classification of depression may be enough in the context of this study. Classifying different levels of depression may be treated as a future study.

Table 2.9 BDI Levels According to the Scores

Level	Score
Minimal	0-14
Mild	14-20
Moderate	20-29
Severe	29-63

STAI is another assessment tool used by psychologists and scholars to measure the existence and severity of state and/or trait anxiety. There are 20 items for State Anxiety

Inventory (SAI) and 20 items for Trait Anxiety Inventory (TAI) part. While SAI items measure the current anxiety, TAI items measure the general anxiety. Higher scores show higher degree of related anxiety type. The scores of SAI and TAI are considered as Anxiety Score together. (See Table 2.10 for some sample items of STAI.)

Table 2.10 Some Example Items of STAI

	Hiç- 1 Biraz- 2 Çok- 3 Tamamen- 4				**Not at all-1 A little-2 Somewhat-3 Very much so-4**			
1 Şu anda sakinim.	1	2	3	4	1 I feel calm	1	2	3	4
2 Kendimi emniyette hissediyorum.	1	2	3	4	2 I feel secure	1	2	3	4
3 Şu anda sinirlerim gergin.	1	2	3	4	3 I feel tense	1	2	3	4
4 Pişmanlık duygusu içindeyim.	1	2	3	4	4 I feel strained	1	2	3	4
5 Şu anda huzur içindeyim.	1	2	3	4	5 I feel at ease	1	2	3	4
6 Şu anda hiç keyfim yok.	1	2	3	4	6 I feel upset	1	2	3	4
7 Başıma geleceklerden endişe ediyorum.	1	2	3	4	7 I am presently worrying over possible misfortunes	1	2	3	4
8 Kendimi dinlenmiş hissediyorum.	1	2	3	4	8 I feel satisfied	1	2	3	4

No inventory is used for the second part of the instances because the instances are already diagnosed by psychologists.

2.5 OTHER SOURCES

2.5.1. Zemberek

Some other utilities are required in the context of this study. Since this study focuses on word usage, it is crucial to obtain root and affixes of each word. For example “my book” can be said as “benim kitabım” or “kitabım” in Turkish. Hence to get the word “benim”, we need to obtain the suffix “ım” from “kitabım”. A morphological analyzer is essential to achieve the analyses similar to the example. There already exists an implemented, publicly available Turkish morphological analyzer called Zemberek (Akın and Akın, 2007), and it is used for accomplishing the task mentioned above. Zemberek analyzes words and provides root-affix combinations with types of words for a given word. However, ambiguity remains as a serious problem, since Turkish has an intrinsic structural ambiguity due to its agglutinative property. Many words present more than one results for morphological analysis. Informatively, to the extent of author's knowledge, there is no successful algorithm solving the ambiguity problem in Turkish. Because the study is not about solving ambiguity and ambiguity is a hard to attack problem to solve, the correct root-affixes combination is chosen manually. Zemberek is an open source program and has Java libraries which can be embedded in the code. Figure 2.2 demonstrates a screen shot of executable version of Zemberek. Table 2.11 presents an example of morphological analysis of a sentence by Zemberek.

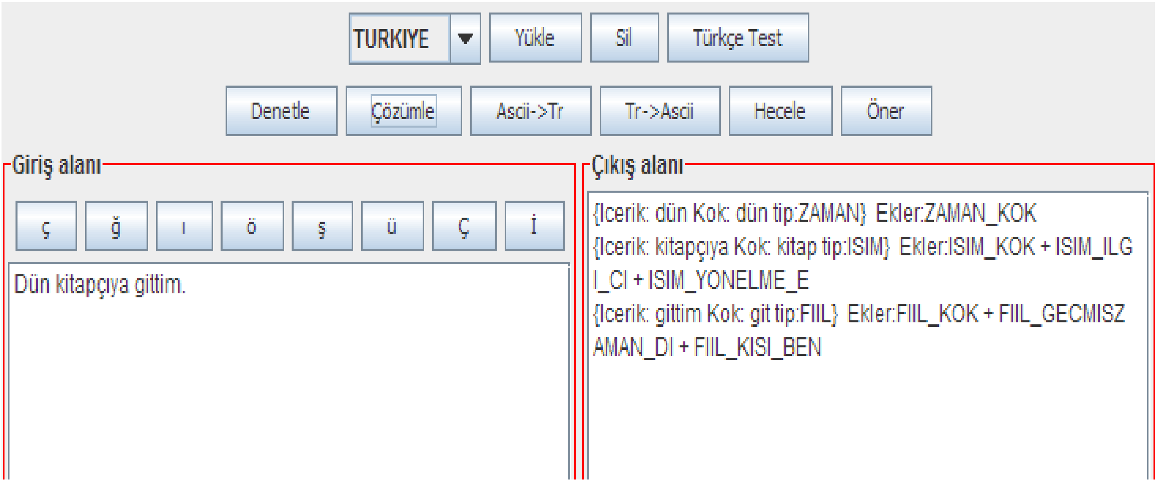

Figure 2.2 Executable Version of Zemberek

Table 2.11 Morphological Analysis Results of a Sentence by Zemberek

Word	Morphological Analysis
Bu	{Icerik: bu Kok: bu tip:ZAMIR} Ekler:ZAMIR_KOK
Gün	{Icerik: gün Kok: gün tip:ZAMAN} Ekler:ZAMAN_KOK
Hava	{Icerik: hava Kok: hava tip:ISIM} Ekler:ISIM_KOK
	{Icerik: hava Kok: hav tip:ISIM} Ekler:ISIM_KOK + ISIM_YONELME_E
	{Icerik: hava Kok: hav tip:ISIM} Ekler:ISIM_KOK + ISIM_YONELME_E
güzel.	{Icerik: güzel Kok: güzel tip:SIFAT} Ekler:ISIM_KOK

2.5.2. Weka

To evaluate the word usage results of people who are diagnosed or not diagnosed with depression and/or anxiety, it is required to execute some classification methods. Data classified according to their diagnosis by considering their word usage analysis results. Then the classification results are examined to detect the effect of psychological disorders on word usage. A data mining tool called Weka (Frenk and Witten, 2005) is used to achieve this goal. Weka is also an open source program and has libraries in Java. It gives the opportunity to use these libraries in various project codes. Weka provides many well known machine learning algorithms as classifiers that can be applied to various data. It has the facility to test and train the classifiers. It gives results of each training and testing process. Weka also permits filtering the attributes (features) or instances of data. Figure 2.3 provides a screen shot of executable version of Weka.

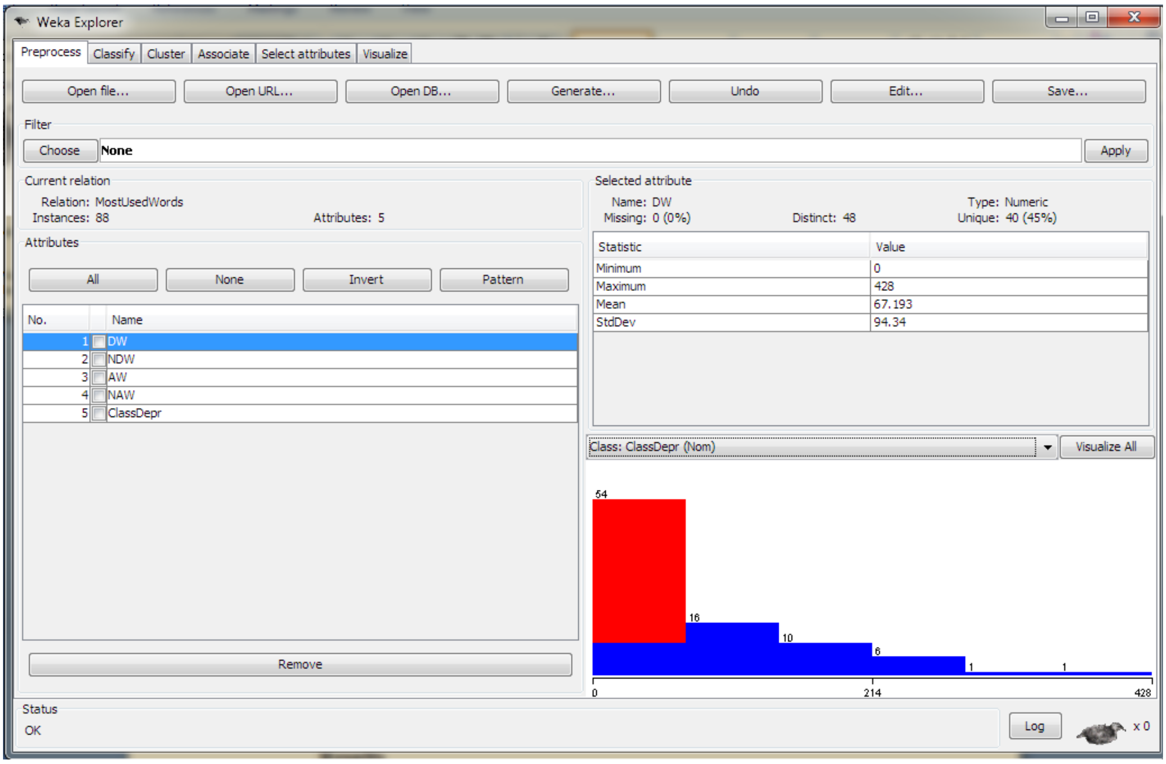

Figure 2.3 A Screenshot of Weka

Weka uses Comma Separated Values (CSV) file format besides many other formats. A file with CSV format involves a data table. Each row of the data table represents a record and each value of the record is separated by comma. CSV formatted datasets are formed to train and test the data mentioned before with Weka. CSV files can be read by Excel. In Table 2.12 some sample values are presented. ID column shows the ID of the instance. Since the names of the patients are hidden, IDs are used to distinguish the instances. NOUN and VERB columns represent the attributes for usage of nouns and verbs. ClassDepr column represents the diagnoses of the instance for depression and ClassAnx represents the diagnoses of the instance for anxiety.

Table 2.12 Some Example Values from the NounVerb CSV File of Depression

ID	**NOUN**	**VERB**	**ClassDepr**
21	0.48	0.30	no
11600	0.47	0.24	yes
58	0.67	0.21	yes
69	0.58	0.18	no
28	0.57	0.26	no
40	0.58	0.20	no
86	0.69	0.24	no
14	0.58	0.24	no
22	0.58	0.24	yes

2.6 TURKISH INVESTIGATOR FOR PSYCHOLOGICAL DISORDERS (TIPD)

TIPD program is developed in Java language and consists of four packages and a Main class. These packages are:

- Data
- DataSet
- Evaluation
- InputOutput

Data package involves classes for handling the data. This package is composed of six classes. These are:

- Info
- Person
- Sentence
- Test
- Text
- Word

DataSet package involves classes to form the datasets to use in classification process. DataSet package has four classes. These are:

- MostUsedWords
- TensePersonNounVerb
- WordsOnly

Evaluation package involves the classes for the classification process. Evaluation package includes one class called TryWeka. InputOutput package involves the classes for the file reading and writing processes. InputOutput package consists of three classes. They are:

- Path
- Reader
- Writer

(See Table 2.13 for the whole structure of the program.)

An example sequence of processes while creating the data of a person is as follows:

- The raw versions of text, info and inventory answers are read from a plain text files.
- A Person object including a Text, an Info and a Test object is formed.

- The raw text is divided into sentences by checking semi columns and a Sentence object for each sentence is formed.
    - Each sentence is divided into words by the Sentence objects and a Word object for each word is formed.
    - Each word is morphologically analyzed by the Word class.
- The Person classes are saved as files to be able to accessed later.

2.6.1. Data Package

The classes in data package are developed in such a manner that it supports easy access and easy modification of the data. The classes in data package, not only handle data, but also provide initializing some values semi automatically while reading data.

2.6.1.1. Person Class

Person class is the main class to handle data of an instance. A Person object includes:

- A Text object
- An Info object
- A Test object
- the id of the instance

The Text object handles the writings of the instance. The Info object is to store some information of the instance. Test object is for handling the test answers and scores. Each instance is stored with a unique id. (See Table 2.14 for the structure of a Person object.)

2.6.1.2. Info Class

Info class is developed to handle some information of the instances, information like birth place, education level etc. Because of different sources of data, this class cannot be used perfectly. The handled information cannot be added to the training and testing process. This class is left because it can be used with new data with complete information. Info class extracts the information from a line of text.

- **Example of a line of information:** Ankara-Kocaeli, Samsun-Master-Bilgisayar Mühendisliği (birth place-other places lived in-education-branch)

2.6.1.1. Test Class

Test class is for handling test answers and scores of the instances. While initializing, it gets the answers of the two tests, BDI and STAI. Then it calculates the scores according to the answers of the tests. A function of this object sets the class values of the instance depending on its scores. The classes can also be changed later on. This property is provided because diagnosed instances have no inventory answer.

Table 2.13 The Classes of the TIPD Program

Package	Classes	Brief info
Data	Info	Class to handle personal info of an instance
	Person	Class to form a person object that includes an info object, a test object and a text object
	Sentence	Class to handle a sentence of an instance
	Test	Class to handle the test answers and results of an instance
	Text	Class to handle the writing of an instance
	Word	Class to handle a word and results of Zemberek analysis of the word
DataSet	MostUsedWords	Class to form the groups of most used words for each class and form a dataset consist of most used words values for each instance
	TensePersonNounVerb	Class to form a dataset with tense, personal pronoun, noun and verb usage values
	WordsOnly	Class to form a dataset with usage values of each word
Evaluation	TryWeka	Class to do the train and test process with the datasets by using 10-fold cross validation
InputOutput	Path	Class to handle the paths of the data
	Reader	Class to use for reading operations
	Writer	Class to use for writing operations

2.6.1.1. Text Class

Text class is to handle the writings of an instance. A Text object includes sentence objects as much as the sentences in the writings of the instance. While a Text object is initializing, it gets the writings of the instance in a raw txt format as a parameter. Then it divides the text into sentences by checking the full stops. Since the sentences are detected by checking full stops, texts are checked manually to remain only the dots at the end of the sentences. For example if second is written as “2.” it is converted into “second”.

2.6.1.1. Sentence Class

Each object of Sentence class handles a sentence of the writings of an instance. Sentence object includes the raw text format of the sentence and an array of Word objects that handles the words in the sentence. Sentence class divides the raw sentence text into words while initializing. The words are extracted considering spaces.

Table 2.14 The Structure of a Person Object

<table>
<tr><td colspan="9">Person</td></tr>
<tr><td rowspan="3">Info</td><td rowspan="3">Test</td><td colspan="7">Text</td></tr>
<tr><td colspan="3">Sentence</td><td colspan="3">Sentence</td><td rowspan="2">…</td></tr>
<tr><td>Word</td><td>Word</td><td>…</td><td>Word</td><td>Word</td><td>…</td></tr>
</table>

2.6.1.2. Word Class

Word class is to handle each word of a sentence. Word objects include the raw text version of the word, an array of strings that consists of the results of the analysis given by Zemberek for the word, a string that handles the Zemberek analysis result chosen by the user, and a string that handles the root of the chosen Zemberek analysis result. While a Word object is being initialized it forms a Zemberek object and analyses the word by using it. Since Zemberek gives multiple analysis results, Word object asks the user to choose one of them. When the word is incorrect, Zemberek gives no results. In this case Word object warns the user about the incorrect word and asks for the correct one. This is the semi automatic part of the software.

2.6.2. InputOutput Package

InputOutput package carries out the file writing and reading processes. The classes of the package are; Reader, Writer and Path. File writing processes are done by using Writer class. It has 2 static methods. These are SavePersonObject and SaveCSV. SavePersonObject is used to save Person objects as a file. SaveCSV method is used to save given String as a CSV formatted file.

2.6.2.1. Reader Class

Reader class has 6 static methods. These are ReadPersonObject, ReadPeople, ReadRawTexts, ReadInfo, ReadTests, and ReadNames. ReadPersonObject method gets an id and reads the Person object with the given id. ReadPeople method reads all Person objects and returns as a HashSet. ReadRawTexts method gets an id as a parameter. It reads the text of the Person with the given id from a txt file. Then forms a Text object and returns it. ReadInfo method gets an id as a parameter. Reads info of the instance with the given id from a txt file and forms an Info abject and returns it. ReadTests method gets an id as a parameter. Then it reads from the test file in txt format. Forms a Test object and returns it. ReadNames method gets a path as a parameter. Reads the names of the files in the path and returns it.

2.6.2.2. Path Class

Path class consists of Strings that shows paths of different files. Reader and Writer classes do their processes considering the paths written in the path object. Paths can be changed by changing only the Path class.

2.6.2.3. Main Class

Main class is the class where Person objects are formed. We can call this process as the beginning of the processes in this study. Main class reads the text files in txt format from the path that it reads from Path class. This is done by Reader class. Then it forms Person objects by giving an id each of them. Each id is one of the names of the text files. Person object reads the text, information and test files with the given id and continues other processes by itself.

2.6.3. DataSet Package

DataSet package consists of classes that form data files to use for train and test process. These classes form CSV files of different attributes to be executed by Weka. These attributes are:

- Mostly used words of each group
- Tenses
- Personal pronouns
- Nouns
- Verbs
- All of the words used by the instances

These CSV files also have class values as attributes. These values are “yes” and “no” for depression and anxiety. Yes means “diagnosed as” for each group of disorder. no means “ not diagnosed as” for each of the disorder group. So there are four groups of instances and some of them intersect. These groups are diagnosed as depression, not diagnosed as depression, diagnosed as anxiety, and not diagnosed as anxiety. Groups about depression and anxiety may intersect. (See Figure 2.4 for the representation of the groups.)

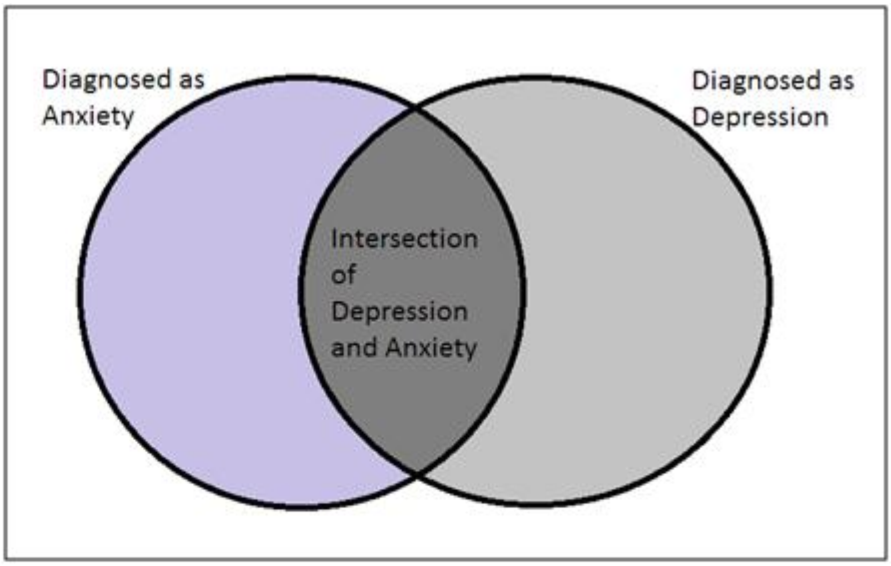

Figure 2.4 Representation of Different Instance Groups

2.6.3.1. MostUsedWords Class

MostUsedWords class explores the words mostly used by each group of instances. First of all, it reads all of the instances by using the ReadPeople method of Reader class. It forms hash tables for each group of instances with the frequency values of word usage number of the group for each word. The groups are called as follows Depressive Words (DW), Non Depressive Words (NDW), Anxious Words (AW), and Non-Anxious Words (NAW). (See Table 2.15 for some examples of most used words and frequency values.)

Table 2.15 Some Examples of Most Used Words and Usage Values by the Instances

DW	Usage	NDW	Usage	AW	Usage	NAW	Usage
çok	0.019	**bir**	0.041	**bir**	0.037	**çok**	0.021
ben	0.014	**ve**	0.029	**ve**	0.029	**yap**	0.011
yap	0.011	**bu**	0.022	**bu**	0.022	**kendi**	0.010
iste	0.009	**insan**	0.011	**insan**	0.010	**daha**	0.010
kendi	0.009	**çalış**	0.010	**için**	0.009	**iş**	0.010

After forming hash tables the class removes the words with close usage values from both of the groups of same disease. The similarity is determined considering a threshold value. Threshold is compared with the difference and sum of the usage values of the word for the groups. The example formula for depression is shown below

UVD=Usage Value in Depression Group

UVND=Usage Value in Non Depression Group

CV=Compared Value

CV=|UVD-UVND|/UVD+UVND

If CV is smaller than threshold, the words will be removed from both of the groups of the same disease. Else the word is removed from the group with smaller value. After the formation of the four groups, the word usage values for each instance are calculated. The values are determined as the frequencies of words of each group for each instance.

A CSV file consists of these values is written to use with the train and test process. (See Table 2.16 for some example values from the MostUsedWords CSV file.)

Table 2.16 Some Example Values from the MostUsedWords CSV File of Depressio

ID	DW	NDW	AW	NAW	ClassDepr
59	0.012	0.980	0.953	0.036	no
7800	0.302	0	0.001	0.301	yes
3908602	0.223	0	0.026	0.196	yes
86	0.008	0.139	0.135	0.010	no
6800	0.112	0	0.022	0.087	yes
21	0.001	0.074	0.059	0.013	no

2.6.3.2. TensePersonNounVerb Class

TensePersonNounVerb class forms a CSV file that includes the usage values of tenses, pronouns, nouns and verbs for each instance. Three CSV files are filled by the class. These are Tense, Person and NounVerb. For each of the processes the first step is reading the person objects by using the ReadPeople method of Reader class.

The attributes of the Tenses CSV file are:

- ID (the id of the instance)
- GENIS (for simple present tense)
- GECMIS_DI (for past tense)
- GECMIS_MIS (for past tense)
- SIMDIKI (for present continuous)
- GELECEK (for future tense)
- ClaassDepression
- ClassAnx

The tense properties of the words are determined considering the information included in the results of Zemberek analyses in each Word object. In the results present simple tense is represented with FIIL_GENISZAMAN_IR, past tense is represented with FIIL_GECMISZAMAN_DI and past perfect tense is represented with FIIL_GECMISZAMAN_MIS, present continuous tense is represented with FIIL_SIMDIKIZAMAN_IYOR, and future tense is represented with FIIL_GELECEKZAMAN_ECEK. These properties of the analysis results are checked to determine the tense of each of the words of each instance.

The frequency of each tense is calculated by counting the Word objects with related tense label for each instance. Then the results are saved into a CSV file called Tenses. (See Table 2.17 for some example values from Tenses CSV file.)

Table 2.17 Some Example Values from Tenses CSV File of Depression

ID	**GENIS**	**GECMIS_DI**	**GECMIS_MIS**	**SIMDIKI**	**GELECEK**	**ClassDepression**
64	0.031	0	0	0.005	0.015	yes
11	0.005	0.008	0.008	0.043	0.002	no
38222	0.122	0.040	0	0	0	yes
8	0.013	0	0	0.060	0.004	no
83	0.006	0.003	0.058	0.025	0.006	yes
22	0.011	0	0.002	0.002	0.005	yes
80	0.029	0.005	0.011	0.053	0.005	no

The attributes of the CSV file called Person are:

- ID (the id of the instance)
- BEN (first person singular)
- SEN (second person singular)
- O (third person singular)
- BIZ (first person plural)
- SIZ (second person plural)
- ONLAR (third person plural)
- BEN_BIZ (first person singular or plural)
- SEN_SIZ (second person singular or plural)
- O_ONLAR (third person singular or plural)
- ClassDepression

- ClassAnx

These attributes are used to obtain the frequency values of pronouns. (See Table 2.18 for some example values from the Person CSV file.)

The tags for pronouns in Zemberek results are; BEN, SEN, O, BIZ, SIZ, and ONLAR. These tags are explored and counted to calculate the frequency values.

Table 2.18 Some Example Values from the Person CSV File of Depression

ID	**BEN**	**SEN**	**O**	**BIZ**	**SIZ**	**ONLAR**	**BEN_ BIZ**	**SEN_ SIZ**	**O_ ONLAR**	**Class Depression**
86	0.078	0	0.156	0.019	0.013	0.006	0.098	0.013	0.163	no
4700	0.182	0	0.086	0.008	0	0	0.191	0	0.086	yes
26	0.089	0.006	0.038	0	0.038	0	0.089	0.044	0.038	no
12	0.057	0.005	0.085	0.020	0	0.005	0.078	0.005	0.090	no
71	0.068	0	0.071	0.006	0.003	0	0.074	0.003	0.071	no
47	0.009	0	0.103	0	0	0	0.009	0	0.103	yes
9700	0.233	0	0.025	0	0.012	0	0.233	0.012	0.025	yes
22	0	0.002	0.130	0	0	0	0	0.002	0.130	yes
58	0.009	0	0.109	0	0	0	0.009	0	0.109	yes
39079	0.166	0	0.025	0.003	0.007	0.010	0.170	0.007	0.036	yes

The attributes of the CSV file called NounVerb are:

- ID (the id of the instance)
- NOUN (for nouns)
- VERB (for verbs)
- ClassDepression
- ClassAnx

The tags in Zemberek analysis results are; ISIM_YALIN_BOS, FIIL_YALIN_BOS, FIIL_DONUSUM_BOS, FIIL_BELIRTME_DIK, and ISIM_DONUSUM_BOS. ISIM_YALIN_BOS tag represents a noun which is a noun with no derivational affix. FIIL_YALIN_BOS tag represents a verb which is a verb with no derivational affix. FIIL_DONUSUM_BOS and FIIL_BELIRTME_DIK represent a word that is became a noun from a verb. ISIM_DONUSUM_BOS represents a word that is transformed into a verb from a noun. These tags are extracted from each Word object of each instance and counted in order to obtain the frequencies. (See Table 2.19 for some example values from the NounVerb file.)

Table 2.19 Some Example Values from NounVerb File of Depression

ID	NOUN	VERB	ClassDepr
9500	0.553	0.214	yes
3	0.593	0.238	no
9100	0.502	0.307	yes
5	0.575	0.245	no
6800	0.541	0.234	yes
86	0.692	0.241	no
62	0.516	0.254	no
14	0.588	0.245	no

2.6.3.3. WordsOnly Class

WordsOnly class forms a CSV file that consists of the frequency value of each word for each instance. First of all it forms a feature vector of words used by the instances. Then it counts usage values of each word for each instance. Finally it calculates the frequency value of each word by dividing the counts to total number of words for each instance. A CSV file called WordsOnly is written with these results.

2.6.4. Evaluation Package

Evaluation package has one class called TryWeka. That class evaluates the data sets formed before. It uses some classes of Weka program. These Weka classes provide the use of many classifiers.

2.6.4.1. TryWeka Class

TryWeka class reads CSV files from two directories, Depression and Anxiety directory. So the CSV files are manually changed and moved to these directories. ClassAnx attribute of the CSV files in Depression directory is removed. Similarly ClassDepr attribute of the CSV files in Anxiety directory removed. Each CSV file that is read from one of these directories is sent to each of the classifiers that are provided by Weka. Results of the classification processes are saved into CSV files for Depression and Anxiety separately. Result of each classification process includes percentage of correctly and incorrectly classified instances, true positive and false positive rates, and precision, recall and F-measure values. (See Table 2.20 for some example results.)

Table 2.20 Results of Some Classifiers Trained with the Tenses Data of Depression

Classifier Name	Correctly Classified	Incorrectly Classified	Precision for no	Precision for yes	Recall for no	Recall for yes	FMeasure for no	FMeasure for yes
SMO	0.82	0.18	0.77	0.91	0.93	0.72	0.84	0.81
NaiveBayes	0.80	0.20	0.76	0.86	0.88	0.72	0.82	0.79
RandomForest	0.79	0.21	0.77	0.82	0.84	0.75	0.80	0.78
BFTree	0.78	0.21	0.75	0.82	0.84	0.72	0.79	0.77
ADTree	0.72	0.28	0.72	0.72	0.72	0.72	0.72	0.72
AdaBoostM1	0.71	0.29	0.67	0.77	0.81	0.61	0.74	0.68

Two Classifiers with the best results are chosen for the study. These are Naïve Bayes and SMO (with their names in Weka).

- **Naive Bayes:** Naive Bayes is a Weka class for a Naive Bayes classifier using estimator classes. The classifier is not an updateable classifier, because the estimator precision values are chosen based on the analysis of the training data. (For more information see the study of George H. John and Pat Langley (John and Langley, 1996).)

- **SMO:** Trains a Support Vector Classifier (SVC) with John Platt’s sequential minimal optimization algorithm. A SVC is a classifier which takes a set of data and for possible two classes predicts the membership of each data according to those classes. (For more information see the study of J. Platt (Platt, 1999))

CHAPTER 3

RESULTS AND COMPARISON

Execution of the TryWeka class is the last step of the application and the study. The class writes the results of the classification process to CSV files. The results are divided into the files according to their attributes and classes. The results are written grouped according to their Weka classifiers to each file.

TryWeka class executes the classifiers of Weka program, as mentioned before. Weka provides many classifiers. Four of the classifiers are chosen. They are chosen according to their correctly classified instance percentages among all classification processes. As mentioned in previous chapter, these classifiers are considered as follows:

- Naive Bayes
- SMO

Table 3.1 shows the percentages of correctly classified depression diagnosed instance numbers for three chosen classifiers and attributes. As mentioned before there are 5 kinds of attributes. Naive Bayes classifiers give result for MostUsedWords, NounVerb, Person, Tenses and WordsOnly as, 0.90, 0.63, 0.71, 0.80, and 0.81, respectively. Results of SMO with the same sequence are, 0.90, 0.62, 0.75, 0.82, and 0.81.

Table 3.1 Percentages of Correctly Classified Depression Diagnosed Instance Numbers for the Chosen Classifiers and the Attributes

Attributes	Naive Bayes	SMO
MostUsedWords	0.93	0.90
NounVerb	0.63	0.62
Person	0.71	0.75
Tenses	0.80	0.82
WordsOnly	0.79	0.75

The best resulting attributes are MostUsedWords attributes. Other attributes are ranged as follows by a narrow margin: Tenses and WordsOnly, Person, and NounVerb.

Table 3.2 shows the percentages of correctly classified anxiety diagnosed instance numbers for the chosen classifiers and attributes. The results of correctly classified anxiety diagnosed instance percentages of Naïve Bayes are 0.84, 0.60, 0.71, 0.71, and 0.68 for MostUsedWords, NounVerb, Person, Tenses, and WordsOnly, respectively. Results for the SMO classifier with the same sequence are, 0.98, 0.53, 0.72, 0.71, and 0.65.

Table 3.2 Percentages of Correctly Classified Anxiety Diagnosed Instance Numbers for Two Chosen Classifiers and the Attributes

Attributes	Naive Bayes	SMO
MostUsedWords	0.87	0.98
NounVerb	0.60	0.53
Person	0.71	0.72
Tenses	0.71	0.71
WordsOnly	0.68	0.65

MostUsedWords is the best kind of attributes with the average of correctly classified anxiety diagnosed instance number percentage result of 0.91. NounVerb gives 0.56, Person gives 0.71, Tenses gives 0.71 and WordsOnly gives 0.66 for the average value.

The results mentioned above are unspecific results of the attributes and the two chosen classifiers. The results for each best attribute are given below.

Evaluation of Tenses and Person attributes are done by using ReliefFAttributeEval (Kira and Rendell, 1992) method within Weka. The method evaluates each attribute by sampling each instance and considering the attribute according to the nearest instances of the same and different classes. ReliefFAttributeEval method chooses GENIS attribute which represents simple present tense among all tenses as the best of the attributes for both depression and anxiety diagnosed instances. (See Table 3.3 and Table 3.5 for the ranking values of the tenses.)

Table 3.3 Tenses and Ranking Values for ReliefFAttributeEval Method with Depression Diagnosed Instances

Tenses	Ranking Values
GENIS	0.03943
SIMDIKI	0.03207
GECMIS_DI	0.01617
GECMIS_MIS	0.01587
GELECEK	0.00723

Table 3.4 The Average Usage Values of Tenses by each Class of Depression

Tense	Depressive	Non-Depressive
GENIS	0.0389	0.0199
SIMDIKI	0.0512	0.0245
GECMIS_DI	0.0226	0.0072
GECMIS_MIS	0.0025	0.0053
GELECEK	0.0019	0.0011

The results show that depression diagnosed people use more simple present and present continuous tense than people who are not diagnosed with depression. Past tense is used more by depression diagnosed people and past perfect tense is used more by people who are not diagnosed with depression. The usage of future tense is very similar among the two classes. (See Table 3.4)

The instances with anxiety use more simple present, present continuous and past tense. Instances that are non-anxious use more past perfect tense. Future tens usage is the same among two classes of anxiety. (See Table 3.6)

Table 3.5 Tenses and Ranking Values for ReliefFAttributeEval Method with Anxiety Diagnosed Instances

Tenses	**Ranking Values**
GENIS	0.01557
SIMDIKI	0.01332
GECMIS_DI	0.00455
GECMIS_MIS	0.00393
GELECEK	0.0029

Table 3.6 The Average Usage Values of Tenses by Each Class of Anxiety

Tense	**Anxious**	**Non-Anxious**
GENIS	0.0365	0.0231
SIMDIKI	0.048	0.029
GECMIS_DI	0.0202	0.0103
GECMIS_MIS	0.0022	0.0054
GELECEK	0.0015	0.0015

ReleiefFAttributeEval method chooses BEN_BIZ attribute which represents the combination of first person singular and first person plural pronouns as the best attribute for depression diagnosed instances. ONLAR attribute is chosen by ReliefFAttributeEval as the best attribute for anxiety diagnosed instances. ONLAR attribute represents third person plural pronoun. (See Table 3.7 and Table 3.9 for personal pronouns and ranking values.)

Table 3.7 Personal Pronouns and Ranking Values for ReliefFAttributeEval Method with Depression Diagnosed Instances

Personal Pronouns	Ranking Values
BEN_BIZ	0.09132
BEN	0.08848
BIZ	0.02776
O_ONLAR	0.02156
O	0.02125
ONLAR	0.01894
SEN_SIZ	0.00996
SEN	0.0059
SIZ	0.00479

Table 3.8 The Average Usage of Personal Pronouns for each Class of Depression

Personal Pronouns	Depressive	Non-Depressive
BEN_BIZ	0.1734	0.0749
BEN	0.1625	0.0627
BIZ	0.0109	0.0123
O_ONLAR	0.0636	0.0887
O	0.057	0.0811
ONLAR	0.0068	0.0077
SEN_SIZ	0.0097	0.0189
SEN	0.0028	0.0055
SIZ	0.007	0.0135

The results obviously show that depression diagnosed people use more first person singular, less third person singular and plural, and less second person singular and plural. (See Table 3.8) This finding confirms that the previous findings for English language about high first person usage of depression diagnosed people are also true for Turkish language. (Rude et al., 2004)

Table 3.9 Personal Pronouns and Ranking Values for ReliefFAttributeEval Method with Anxiety Diagnosed Instances

Personal Pronouns	Ranking Values
ONLAR	0.03507
BEN_BIZ	0.02604
BEN	0.02418
O_ONLAR	0.02376
O	0.01923
SEN_SIZ	0.01908
SIZ	0.01484
SEN	0.01173
BIZ	0.0096

Table 3.10 The Average Usage of Personal Pronouns for each Class of Anxiety

Personal Pronouns	Anxious	Non-Anxious
BEN_BIZ	0.0871	0.1666
BEN	0.0762	0.1543
BIZ	0.011	0.0123
O_ONLAR	0.0873	0.0634
O	0.0785	0.0582
ONLAR	0.0091	0.0052
SEN_SIZ	0.014	0.0147
SEN	0.0045	0.0037
SIZ	0.0097	0.0108

Table 3.10 shows the average usage values of personal pronouns by each class of anxiety. Anxious people use more third person singular and plural, and less first person pronouns.

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 EVALUATION OF RESULTS

A psychological text analysis system called TIPD is represented in this study.

Two utilities that are called Zemberek and Weka are used in order to implement the system. The system uses ML methods provided by Weka to maintain the text analysis process. The analyses are done by using the classification methods provided by Weka.

The classes of the classification process are considered in four groups. These are depressed, non-depressed, anxious, and non-anxious. The text analyses are done for the Turkish language. The data for the classification process is gathered in two ways. One of them is by using a form to fill and the other is by the contributions of the psychology professionals.

The results of the study confirm that language usage is one of the important indicators of psychological state. This study is the first study that maintains Turkish psychological text analysis by using NLP methods, to the knowledge of the author.

4.2 FUTURE IMPROVEMENTS

That kind of studies may be maintained by using many different kinds of data. Analyses can be done by using writings or discourses of different groups like students, politicians, women, men, etc. Different kinds of texts may be used like columns, articles, novels and so on. Different disorders may be analyzed and many new findings may be obtained from these analyses.

There are many error sources for this kind of NLP studies. For this study, ambiguity in morphologic analysis is the biggest problem. Turkish is an agglutinative language and this causes many ambiguity problems because of the possibility of existence of many affixes in a word. It is required to negotiate this handicap in order to advance in this kind of NLP studies.

There are many possible applications for the interdisciplinary studies of NLP and social sciences. It can be said that these applications are necessary for many improvements in social sciences like psychology. Creation of thesauruses or dictionaries by professionals of social sciences is very important to maintain this kind of studies.

As an initial study for Turkish language and psychological text analysis, this study is just the beginning of many other studies that can be maintained in the wide scope of the area. There is still a long way to go for Turkish text analysis to be helpful and contribute to psychology or other social sciences.

REFERENCES

Abbasi, A., and Chen, H., "Affect Intensity Analysis of Dark Web Forums", *Intelligence and Security Informatics 2007*, New Brunswick, May 2007, pp 282-288.

Akın, A. A., and Akın, M. D., *Zemberek, an open source NLP framework for Turkic languages*, 2007, http://zemberek.googlecode.com

Atkinson, J. W., and McClelland, D. C., "The Effect of Different Intensities of the Hunger Drive on Thematic Apperception", *Journal of Experimental Psychology*, Vol. 38, pp 643–658, 1948.

Beck, A. T., Steer, R. A., and Brown G. K., *Beck Depression Inventory Second Edition Manual*, The Psychological Corporation, TX, 1996.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J., "An inventory for measuring depression", *Arch Gen Psychiatry*, Vol 4, pp 61-71, 1961.

Breck, E., Choi, Y., and Cardie, C., "Identifying expressions of opinion in context", *Proceedings of the 20th international joint conference on Artifical intelligence*, India, 2007.

Carbonell, J. G., *Subjective Understanding: Computer Models of Belief Systems*, Doctoral Dissertation, Yale, 1979.

Carver, C. S., and White, T. L., "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales", *Journal of Personality and Social Psychology*, Vol 67, pp 319-333, 1994.

Chomsky, N. A., *Syntactic Structures*, Mouton, The Hauge, 1957.

Chung C., and Pennebaker J. W., "The psychological function of function words", in K. Fiedler (Ed.), *Social Communication*, pp 343-359, Psychology Press, New York, 2007.

Cohen, A. S., Minor, K. S., Baillie, L. E., and Dahir, A. M., “Clarifying the Linguistic Signature: Measuring Personality from Natural Speech”, *Journal of Personality Assessment*, Vol 90, No 6, pp 559 — 563, 2008.

Ding, X., Liu, B., and Yu, P. S., “A Holistic Lexicon Based Approach to Opinion Mining”, *First ACM International Conference on Web Search and Data Mining*, California, February 2008, pp 231-240.

Duda, R., and Hart, P., *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

Fox, S., Zickuhr, K., and Smith, A., *19% of Internet Users on Twitter or Another Service*, 2009, http://pewresearch.org/pubs/1385/who-uses-twitter-tweets

Freud, S., *Psychopathology of Everyday Life*, Basic Books, New York, 1901.

Fukuhara, T., Nakagawa, H., and Nishida, T., "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events", *Proceedings of the International Conference on Weblogs and Social Media*, Colorado, 2007.

Ghose, A., and Ipeirotis, P. G., “Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews”, *Proceedings of the Ninth International Conference on Electronic Commerce*, Minnesota, August 2007, pp 303-310.

Gottschalk, L. A., and Gleser, G. C., *The measurement of psychological states through the content analysis of verbal behavior*, University of California Press, Berkeley, 1969.

Hu, M., Sun, A., and Lim, E. P., “Comments Oriented Blog Summarization by Sentence Extraction”, *Proceedings of the Conference on Information and Knowledge Manegement*, Portugal, November 2007, pp 901-904.

Infogistics, *NLProcessor - Text Analysis Toolkit*, 2000, http://www.infogistics.com/textanalysis.html

Jin, X., Li, Y., Mah, T., and Tong, J., "Sensitive Webpage Classification for Content Advertising", *Proceedings of the First International Workshop on Data Mining and Audience Intelligence for Advertising*, California, August 2007, pp 28-33.

Jindal, N., and Liu, B., "Opinion and Spam Analysis", *Proceedings of the International Conference on Web Search and Data Mining*, California, February, 2008, pp 219-230.

Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relecant Features", *Proceedings of the 10th European Conference on Machine Learning*, Germany, April, 1998, pp 137-142.

John, G. H., and Langley, P., "Estimating Continuous Distributions in Bayesian Classifiers", *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 1996, pp 338-345.

Kim, S. M., and Hovy, E., Crystal: Analyzing Predictive Opinions on the Web, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 2007, pp 1056-1064.

Kira, K., Rendell., L. A., "A Practical Approach to Feature Selection", *Ninth International Conference on Machine Learning*, 171-182, 1994.

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., and Zhou, M., "Low Quality Product Review Detection in Opinion Summarization", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, June 2007, pp 334-342.

McCallum, A. K., MALLET: *A machine learning for language toolkit*, 2002, http://mallet.cs.umass.edu

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

Pennebaker, J. W., "Current issues and new directions in Psychology and Health: Listening to what people say – the value of narrative and computational linguistics in health psychology", *Psychology and Health,* Vol 22, No 6, pp 631-635, 2007.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J., *The Development and Psychometric Properties of LIWC2007,* 2007, http://liwc.net

Pennebaker, J. W., Francis, M. E., and Booth, R. C., *Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program (2nd ed.)*, Erlbaum, NJ, 2001.

Platt, J., "Fast training of support vector machines using sequential minimal optimization", in Schölkopf, B., Burges, C., and Smola, A. (Eds.), *Advances in Kernel Methods - Support Vector Learning*, pp 185-208, MIT Press, 1999.

Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., and Olmstead, K., *Understanding the Participatory News Consumer*, 2010, http://pewresearch.org/pubs/1508/internet-cell-phone-users-news-social-experience

Rude, S. S., Gortner, E. M., and Pennebaker, J. W., "Language Use of Depressed and Depression Vulnerable College Students", *Cognition and Emotion,* Vol 18, No 8, pp 1121-1133, 2004.

Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Computing Surveys,* Vol. 34, No. 1, pp 1-47, 2002.

Soderland, S., "Learning Information Extraction Rules for Semi-Structured and Free Text", *Machine Learning,* Vol. 34, No. 1-3, pp 233-277, 1999.

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E., *Manual for the State Trait Anxiety Inventory*, Consulting Psychologists Press, CA, 1970.

Turing, A. M., "Computing machinery and intelligence", Mind, 59, pp 433-460, 1950.

Watson, D., and Clark, L. A., *PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form,* The University of Iowa, 1999.

Wiebe, J., Wilson, T., and Cardie, C., "Annotating expressions of opinions and emotions in language", *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39 (2-3), pp 165-210, 2005.

Wilks, Y., and Bien, J., "Beliefs, points of view, and multiple environments", *Cognitive Science*, 7, pp 95-119, 1983.

Wilson T., "Annotating subjective content in meetings", *Proceedings of LREC*, 2008.

Winograd, T., *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*, MIT AI Technical Report 235, February 1971.

Witten, I. H., and Frenk E., *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.

Yang, Y., and Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, July 1997, pp 83-91.

Zimmerman, M., and Coryell, W., "The Inventory to Diagnose Depression, Lifetime Version", *Acta Psychiatrical Scandinavia*, Vol 75, pp 495-499, 1987.

APPENDIX A

DATA GATHERING FORM

YÜKSEK LİSANS TEZİ YAZI VE TEST ÇALIŞMASI

Yrd. Doç. Zeynep ORHAN

Araş. Gör. Nur Banu ALBAYRAK

BİLGİLENDİRME

- Bu yazı ve test çalışması "OPINION AND SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING TECHNIQUES" başlıklı yüksek lisans tez çalışması ile ilgili olarak, depresyon ve anksiyete düzeylerinin kestiriminde kullanılmak amacıyla yapılmaktadır.
- Çalışma, yazının psikolojik açıdan analizinin Türkçe ile yapılması için önemlidir.
- Yazı ve test çalışması, belirtilen konularda yazı yazılması ve test cevaplama olarak iki bölümden oluşmaktadır.
- Yazı yazılması işleminin testlerden önce yapılması çalışmanın sağlığı için önemlidir, **lütfen önce yazı yazma bölümünü tamamlayınız**.
- Vereceğiniz bilgiler kesinlikle gizli tutulacak ve psikolojik testler ile yazı analizlerinin **bireysel sonuçları** hiçbir kişi ya da kuruma **verilmeyecektir**.
- Bu çalışmaya katılarak bilimsel bir çalışmaya önemli bir destek vereceksiniz.

Bu çalışmaya göstereceğiniz ilgiden ve ayıracağınız zamandan dolayı şimdiden teşekkür ederiz.

Saygılarımızla.

<u>ÖNEMLİ NOT:</u> Çalışmaya yapacağı katkı için aşağıdaki alanları da doldurmanızı rica ediyoruz. **Eğer bu bilgileri vermek istemiyorsanız, boş bırakabilirsiniz.**

<u>KİŞİSEL BİLGİLER:</u>

Doğum yeri:

Yaşadığınız diğer yerler(Uzun süreli olarak):

Eğitim seviyesi:

Branşı:

BÖLÜM 1: BELİRTİLEN KONUDAKİ DÜŞÜNCELERİ YAZILI İFADE ETME

NOT:

- Verilen konu sayısı 4'tür.
- İlk konu hakkındaki düşüncelerinizi yazı ile ifade etmenizi rica ediyoruz.
- Diğer 3 konu ise isteğe göre yazılabilir veya boş bırakılabilir.
- Ne kadar çok konuda yazı yazılırsa çalışmaya o kadar çok fayda sağlanmış olacaktır.
- Yazılacak yazıların **150-200 kelime** civarında (**yaklaşık 1 sayfa**) olması çalışmada doğru sonuçlara ulaşılabilmesi için önemlidir.

SORU 1 :

Akademisyenlik mesleği, yoğunluğu, daha çok fedakarlık gerektirmesi, bunlarla birlikte toplum içinde daha iyi bir statü kazandırması açısından diğer mesleklerden ayrılmaktadır. Akademisyenlik ile ilgili düşüncelerinizi, tecrübelerinizden de bahsederek anlatınız.

SORU 2 :

Hayatınız boyunca yaşadığınız, sizi **en çok dehşete düşüren** ya da **üzen olayı** ve size düşündürdüklerini anlatınız.

SORU 3 :

Dünyanın durumu ve kendinizin, ailenizin, arkadaşlarınızın bu durumdaki yeri hakkındaki görüşlerinizi yazınız.

SORU 4 :

Geçmişinizi ve bugününüzü karşılaştırarak hayatınızı çeşitli yönlerden (duygular, madde, insan ilişkileri, meslek, sağlık vb.) değerlendiriniz.

BÖLÜM 2: TESTLER

NOT: Testler 21 ve 40 soru olmak üzere 2 tanedir. **Bir test bitirilmeden diğerine geçilmemelidir**.

TEST 1:

AÇIKLAMA: Aşağıda gruplar halinde cümleler verilmektedir. Öncelikle her gruptaki cümleleri dikkatle okuyarak, **BUGÜN DAHİL GEÇEN HAFTA** içinde kendinizi nasıl hissettiğinizi en iyi anlatan cümleyi seçiniz. Sorulara vereceğiniz samimi ve dürüst cevaplar araştırmanın bilimsel niteliği açısından son derece önemlidir.

1- Lütfen en uygun cümleyi işaretleyiniz.
   a. Kendimi üzüntülü ve sıkıntılı hissetmiyorum.
   b. Kendimi üzüntülü ve sıkıntılı hissediyorum.
   c. Hep üzüntülü ve sıkıntılıyım. Bundan kurtulamıyorum.
   d. O kadar üzüntülü ve sıkıntılıyım ki artık dayanamıyorum.
2- Lütfen en uygun cümleyi işaretleyiniz.
   a. Gelecek hakkında mutsuz ve karamsar değilim.
   b. Gelecek hakkında karamsarım.
   c. Gelecekten beklediğim hiçbir şey yok.
   d. Geleceğim hakkında umutsuzum ve sanki hiçbir şey düzelmeyecekmiş gibi geliyor.
3- Lütfen en uygun cümleyi işaretleyiniz.
   a. Kendimi başarısız bir insan olarak görmüyorum.
   b. Çevremdeki birçok kişiden daha çok başarısızlıklarım olmuş gibi hissediyorum.
   c. Geçmişe baktığımda başarısızlıklarla dolu olduğunu görüyorum.
   d. Kendimi tümüyle başarısız biri olarak görüyorum.
4- Lütfen en uygun cümleyi işaretleyiniz.
   a. Birçok şeyden eskisi kadar zevk alıyorum.
   b. Eskiden olduğu gibi her şeyden hoşlanmıyorum.
   c. Artık hiçbir şey bana tam anlamıyla zevk vermiyor.
   d. Her şeyden sıkılıyorum.
5- Lütfen en uygun cümleyi işaretleyiniz.
   a. Kendimi herhangi bir şekilde suçlu hissetmiyorum.
   b. Kendimi zaman zaman suçlu hissediyorum.
   c. Çoğu zaman kendimi suçlu hissediyorum.
   d. Kendimi her zaman suçlu hissediyorum.
6- Lütfen en uygun cümleyi işaretleyiniz.
   a. Bana cezalandırılmışım gibi gelmiyor.
   b. Cezalandırılabileceğimi hissediyorum.
   c. Cezalandırılmayı bekliyorum.
   d. Cezalandırıldığımı hissediyorum.
7- Lütfen en uygun cümleyi işaretleyiniz.

a. Kendimden memnunum.
b. Kendimden pek memnun değilim.
c. Kendime çok kızıyorum.
d. Kendimden nefret ediyorum.

8- Lütfen en uygun cümleyi işaretleyiniz.
a. Başkalarından daha kötü olduğumu sanmıyorum.
b. Zayıf yanlarım veya hatalarım için kendi kendimi eleştiririm.
c. Hatalarımdan dolayı ve her zaman kendimi kabahatli bulurum.
d. Her aksilik karşısında kendimi hatalı bulurum.

9- Lütfen en uygun cümleyi işaretleyiniz.
a. Kendimi öldürmek gibi düşüncelerim yok.
b. Zaman zaman kendimi öldürmeyi düşündüğüm olur. Fakat yapmıyorum.
c. Kendimi öldürmek isterdim.
d. Fırsatını bulsam kendimi öldürürdüm.

10- Lütfen en uygun cümleyi işaretleyiniz.
a. Her zamankinden fazla içimden ağlamak gelmiyor.
b. Zaman zaman içindem ağlamak geliyor.
c. Çoğu zaman ağlıyorum.
d. Eskiden ağlayabilirdim şimdi istesem de ağlayamıyorum.

11- Lütfen en uygun cümleyi işaretleyiniz.
a. Şimdi her zaman olduğumdan daha sinirli değilim.
b. Eskisine kıyasla daha kolay kızıyor ya da sinirleniyorum.
c. Şimdi hep sinirliyim.
d. Bir zamanlar beni sinirlendiren şeyler şimdi hiç sinirlendirmiyor.

12- Lütfen en uygun cümleyi işaretleyiniz.
a. Başkaları ile görüşmek, konuşmak isteğimi kaybetmedim.
b. Başkaları ile eskiden daha az konuşmak, görüşmek istiyorum.
c. Başkaları ile konuşma ve görüşme isteğimi kaybettim.
d. Hiç kimseyle konuşmak görüşmek istemiyorum.

13- Lütfen en uygun cümleyi işaretleyiniz.
a. Eskiden olduğu gibi kolay karar verebiliyorum.
b. Eskiden olduğu kadar kolay karar veremiyorum.
c. Karar verirken eskisine kıyasla çok güçlük çekiyorum.
d. Artık hiç karar veremiyorum.

14- Lütfen en uygun cümleyi işaretleyiniz.
a. Aynada kendime baktığımda değişiklik görmüyorum.
b. Daha yaşlanmış ve çirkinleşmişim gibi geliyor.
c. Görünüşümün çok değiştiğini ve çirkinleştiğimi hissediyorum.
d. Kendimi çok çirkin buluyorum.

15- Lütfen en uygun cümleyi işaretleyiniz.
a. Eskisi kadar iyi çalışabiliyorum.
b. Bir şeyler yapabilmek için gayret göstermem gerekiyor.
c. Herhangi bir şeyi yapabilmek için kendimi çok zorlamam gerekiyor.
d. Hiçbir şey yapamıyorum.

16- Lütfen en uygun cümleyi işaretleyiniz.
a. Her zamanki gibi iyi uyuyabiliyorum.
b. Eskiden olduğu gibi iyi uyuyamıyorum.

c. Her zamankinden 1-2 saat daha erken uyanıyorum ve tekrar uyuyamıyorum.
d. Her zamankinden çok daha erken uyanıyor ve tekrar uyuyamıyorum.

17- Lütfen en uygun cümleyi işaretleyiniz.

a. Her zamankinden daha çabuk yorulmuyorum.
b. Her zamankinden daha çabuk yoruluyorum.
c. Yaptığım her şey beni yoruyor.
d. Kendimi hemen hiçbir şey yapamayacak kadar yorgun hissediyorum.

18- Lütfen en uygun cümleyi işaretleyiniz.

a. İştahım her zamanki gibi.
b. İştahım her zamanki kadar iyi değil.
c. İştahım çok azaldı.
d. Artık hiç iştahım yok.

19- Lütfen en uygun cümleyi işaretleyiniz.

a. Son zamanlarda kilo vermedim.
b. İki kilodan fazla kilo verdim.
c. Dört kilodan fazla kilo verdim.
d. Altı kilodan fazla kilo vermeye çalışıyorum.

20- Lütfen en uygun cümleyi işaretleyiniz.

a. Sağlığım beni fazla endişelendirmiyor.
b. Ağrı, sancı, mide bozukluğu veya kabızlık gibi rahatsızlıklar beni endişelendirmiyor.
c. Sağlığım beni endişelendirdiği için başka şeyleri düşünmek zorlaşıyor.
d. Sağlığım hakkında o kadar endişeliyim ki başka hiçbir şey düşünemiyorum.

21- Lütfen en uygun cümleyi işaretleyiniz.

a. Son zamanlarda cinsel konulara olan ilgimde bir değişme fark etmedim.
b. Cinsel konularla eskisinden daha az ilgiliyim.
c. Cinsel konularla şimdi çok daha az ilgiliyim.
d. Cinsel konular olan ilgimi tamamen kaybettim.

TEST 2:

<u>AÇIKLAMA:</u> Aşağıda kişilerin kendilerine ait duygularını anlatmada kullandıkları birtakım ifadeler verilmştir. Her ifadeyi okuyup, nasıl hissettiğinizi ifadelerin sağ tarafında bulunan kutucuklardan uygun olanını işaretleyerek belirtiniz. Doğru ya da yanlış cevap yoktur. Herhangi bir ifadenin üzerinde fazla zaman kaybetmeksizin anında nasıl hissettiğinizi gösteren cevabı işaretleyiniz.

		Hiç	Biraz	Çok	Tamamen
1-	Şu anda sakinim.	☐	☐	☐	☐
2-	Kendimi emniyette hissediyorum.	☐	☐	☐	☐
3-	Şu anda sinirlerim gergin.	☐	☐	☐	☐
4-	Pişmanlık duygusu içindeyim.	☐	☐	☐	☐
5-	Şu anda huzur içindeyim.	☐	☐	☐	☐
6-	Şu anda hiç keyfim yok.	☐	☐	☐	☐
7-	Başıma geleceklerden endişe ediyorum.	☐	☐	☐	☐
8-	Kendimi dinlenmiş hissediyorum.	☐	☐	☐	☐
9-	Şu anda kaygılıyım.	☐	☐	☐	☐
10-	Kendimi rahat hissediyorum.	☐	☐	☐	☐
11-	Kendime güvenim var.	☐	☐	☐	☐
12-	Şu anda asabım bozuk.	☐	☐	☐	☐
13-	Çok sinirliyim.	☐	☐	☐	☐
14-	Sinirlerimin çok gergin olduğunu hissediyorum.	☐	☐	☐	☐
15-	Kendimi rahatlamış hissediyorum.	☐	☐	☐	☐
16-	Şu anda halimden memnunum.	☐	☐	☐	☐
17-	Şu anda endişeliyim.	☐	☐	☐	☐
18-	Heyecandan kendimi şaşkına dönmüş hissediyorum.	☐	☐	☐	☐
19-	Şu anda sevinçliyim.	☐	☐	☐	☐
20-	Şu anda keyfim yerinde.	☐	☐	☐	☐
21-	Genellikle keyfim yerindedir.	☐	☐	☐	☐
22-	Genellikle çabuk yoruluyorum.	☐	☐	☐	☐
23-	Genellikle kolay ağlarım.	☐	☐	☐	☐
24-	Başkaları kadar mutlu olmak isterim.	☐	☐	☐	☐
25-	Çabuk karar veremediğim için fırsatları kaçırırım.	☐	☐	☐	☐
26-	Kendimi dinlenmiş hissederim.	☐	☐	☐	☐
27-	Genellikle sakin, kendime hakim ve soğukkanlıyım.	☐	☐	☐	☐
28-	Güçlüklerin yenemeyeceğim kadar biriktiğini hissederim.	☐	☐	☐	☐
29-	Önemsiz şeyler hakkında endişelenirim.	☐	☐	☐	☐
30-	Genellikle mutluyum.	☐	☐	☐	☐
31-	Her şeyi ciddiye alır ve etkilenirim.	☐	☐	☐	☐
32-	Genellikle kendime güvenim yoktur.	☐	☐	☐	☐
33-	Genellikle kendimi güvende hissederim.	☐	☐	☐	☐
34-	Sıkıntılı ve güç durumlarla karşılaşmaktan kaçınırım.	☐	☐	☐	☐
35-	Genellikle kendimi hüzünlü hissederim.	☐	☐	☐	☐
36-	Genellikle hayatımdan memnunum.	☐	☐	☐	☐
37-	Olur olmaz düşünceler beni rahatsız eder.	☐	☐	☐	☐
38-	Hayal kırıklıklarını öylesine ciddiye alırım ki hiç unutamam.	☐	☐	☐	☐
39-	Aklı başında ve kararlı bir insanım.	☐	☐	☐	☐
40-	Son zamanlarda kafama takılan konular beni tedirgin eder.	☐	☐	☐	☐