# SENTIMENT ANALYSIS AND GENDER PREDICTION IN TWITTER DATA

**ERTUĞRUL BALABAN**

**AUGUST 2015**

# SENTIMENT ANALYSIS AND GENDER PREDICTION IN TWITTER DATA

**A THESIS SUBMITTED TO**

**THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF**

**ÇANKAYA UNIVERSITY**

**BY**

**ERTUĞRUL BALABAN**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF**

**MASTER OF SCIENCE**

**IN**

**THE DEPARTMENT OF**

**COMPUTER ENGINEERING**

**AUGUST 2015**

Title of the Thesis: **Sentiment Analysis and Gender Prediction in Twitter Data**

Submitted by **Ertuğrul BALABAN**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof. Dr. Taner ALTUNOK
Director

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

Prof. Dr. Müslim BOZYİĞİT
Head of Department

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

Yrd. Doç. Dr. Abdül Kadir GÖRÜR
Supervisor

**Examination Date: 13.08.2015**

**Examining Committee Members**

| | | |
|---|---|---|
| Assist.Prof. Dr. Abdül Kadir GÖRÜR | (Çankaya Univ.) | |
| Assist. Prof. Dr. Reza HASSANPOUR | (Çankaya Univ.) | |
| Assoc. Prof. Dr. Fahd JARAD | (THK Univ.) | |

## STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Ertuğrul BALABAN

Signature :

Date : 13.08.2015

# ABSTRACT

## SENTIMENT ANALYSIS AND GENDER PRECITION IN TWITTER DATA



BALABAN, Ertuğrul

M.Sc., Department of Computer Engineering

Supervisor: Yrd. Doç. Dr. Abdül Kadir GÖRÜR



August 2015, 39 pages



In this thesis, tweets from Twitter that have been sent by users will be considered on a preferential basis in accordance with determined or requested specific key word(s). Also the interpretation of these tweets, by the computer, will be examined in a way as they are "Positive", "Negative" or "Neutral". In this context, under the heading 'Twitter Sentiment Analysis', studies were conducted and the success rates of achieved results were compared. In addition to this, on the basis of the usernames(of users) who send tweets, tweets was compared with Turkish Special Names which is shared by the Turkish Language Association (TDK) and also achieved results and gender determinations of users in terms of "Female", "Male" or "Not Determined," were examined. Under the heading of 'Gender Prediction in Twitter' studies were conducted and the success rates of achieved results were compared. On the basis of this study, the related topics of ‘Sentiment Analysis' and 'Gender Prediction' were examined for Turkish Language and all of these studies were carried out through Turkish language.



**Keywords:** Sentiment Analysis, Twitter, Gender Prediction in Twitter.

# ÖZ

## TWITTER VERİLERİNDE DUYGU ANALİZİ VE CİNSİYET TESPİTİ

BALABAN, Ertuğrul
Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı
Tez Yöneticisi: Yrd. Doç. Dr. Abdül Kadir GÖRÜR
Ağustos 2015, 39 sayfa

Bu tezde, sosyal medyada önemli bir yere sahip olan Twitter uygulamsında, belirlenen veya istenen spesifik bir kelime (ler)' e göre kişilerin atmış olduğu tweetler ele alınmış, bu tweetlerin otomatik olarak "Olumlu", "Olumsuz" ya da "Nötr" şeklinde bilgisayar tarafından yorumlanması incelenmiştir. Bu kapsamda, "Twitter Duygu Analizi" başlığı altında çalışmalar yapılmış ve çıkan sonuçların başarı oranları karşılaştırılmıştır. Bunun yanı sıra, tweet atan kullanıcıların kullanıcı adları temel alınarak, Türk Dil Kurumu tarafından paylaşılan Türkçe Özel İsimler ile karşılaştırılmış, çıkan sonuçlar ve eşleşmelere göre kullanıcıların Cinsiyet tespitlerinin "Kadın", "Erkek" ya da "Tespit Edilemedi" şeklinde yorumlanması incelenmiştir. Yine bu kapsamda da, "Twitterda Cinsiyet Tespiti" başlığı altında çalışmalar yapılmış ve çıkan sonuçların başarı oranları karşılaştırılmıştır. Bu çalışmanın temelinde, Twitter Duygu Analizi ve Twitterda Cinsiyet Tespiti başlılklı konular Türkçe dili için ele alınmış ve yapılan tüm çalışmalar Türkçe dili üzerinden yürütülmüştür.

**Anahtar Kelimeler:** Duygu Analizi, Twitter, Twitterda Cinsiyet Tespiti.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Yrd. Doç. Dr. Abdül Kadir GÖRÜR for his supervision, special guidance, suggestions, and encouragement through the development of this thesis.

It is a pleasure to express my special thanks to my family and friends for their valuable support.

**TABLE OF CONTENTS**

## LIST OF FIGURES

### FIGURES

## LIST OF TABLES

### TABLES

**TABLES**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| FB | Fenerbahçe |
| FK | Fleiss’ Kappa |
| GD | Gender Detection |
| GS | Galatasaray |
| KBSR | Knowledge Based Semantic Relatedness |
| NLP | Natural Language Processing |
| PC | Pearson Correlation |
| RP | Recall and Precision |
| SA | Sentiment Analysis |
| SR | Semantic Relatedness |
| SS | Semantic Similarity |
| SWW | Senti Wordnet Weights |
| VR | Value Range |

# CHAPTER 1

# INTRODUCTION

## 1.1 Sentiment Analysis and Gender Prediction

Sentiment analysis is the automatic classification method to identify and extract subjective information from words, phrases or text. It used to detect positive or negative opinions which are specified by documents that were examined by computer. Analyzing each word manually is almost impossible task because of a big data and time related factors for comments, we need the tool which can make sentiment analysis automatically. This process can be used especially in social networks as Twitter to analyze user opinions. Because a lot of companies, politicians or famous people have their own official Twitter accounts. Thus, making automatically sentiment analysis ensures monitoring opinions and thoughts of users about them. The consumers who want to investigate the product sentiments preparatory to buying can benefit from this, as well as the companies which want to track their trademarks' public sentiment.

Statistical approach and Linguistic approach are the main two approaches used in sentiment analysis[1]. Statistical approach depends on comparing the negative and positive statement amount statistically and mathematically. However in the text and verbal approach, the analyzed text analysis besides rule build are the processes operated.

Additionally, gender prediction has an important part of text mining, that gave us an additional information about the author who made the comments. In Twitter, username and profile picture can be helpful to detect the gender of the comment and analyzing the percentage of the woman or man thoughts. Thus, it can be used in different applications across a variety of fields like marketing, advertising and legal inverstigation.

### 1.2 Objectives

The main objective of this research is to interpret the text comments (i.e. tweets) by the computer, which are collected from Twitter social network then examined in a way as they are "Positive", "Negative" or "Neutral". While deciding on comments of tweets, gender prediction of tweets is the second main task for this thesis.

In general, tweets from Twitter send by user about the certain issue or related with the selected key words, provides to show the thoughts of the people about it. This thesis provides to explain the emotions of the tweets and gender of the person who send the tweets. Because of its popularity and commercial revenue, twitter tweets preferred to interpret. Besides, men and women think differently even though they talk technically same language. Gender of tweets show this difference and give a percentage of the results on comments as women or men.

The keywords and sentences can be used to search this big quantity. However, the meaning and the sentiment meaning of the text must be realized by search engines for consequence achievement. The sentiment analysis is the most proper solution in order to solve this issue. It has key concern which is the foundation of the feelings that the text gives.

Unfortunately, all studies in the literature about Sentiment analysis generally focused on the English language. There is no enough Turkish study or tool to implement this work. Therefore, in this thesis, generate Turkish data set and provide Turkish Sentiment Analysis is the main purpose. Besides, because of Turkish data, people who share these data (tweets) on their own accounts also critical task to determine the gender of the tweets. Therefore, there should be also Turkish Name data set creation to decide the gender of the person who sent tweets from Twitter. There is no any kind of study in Turkish also.

According to studies above, the main idea in this thesis is to apply Turkish Sentiment Analysis on the Twitter data and also provide the gender prediction for the Turkish tweets. This thesis provides new methods for the structure of Sentiment Analysis and Gender Prediction for the Twitter data.

## 1.3 Outlines

This thesis contains five chapters. All the necessary information about the Turkish Sentiment Analysis and Gender Prediction from the Twitter data will be examined. According to created Turkish data set, comparison with the other studies and successful rates show future of this study. According to the main concept of the thesis;

Chapter 1 is an introduction to the sentiment analysis and gender prediction on Twitter data and objectives of this thesis.

Chapter 2 includes background information related with the implementation studies and proper application used in.

In Chapter 3, methods will be explained to apply sentiment analysis on Turkish Twitter data. Besides methods for gender prediction on Turkish Twitter data.

In Chapter 4, according to implementation technic which were used in this thesis will show the success rate results and provide to compare other studies done in same way.

Chapter 5 includes the conclusion and future work part.

# CHAPTER 2

# BACKGROUND FOUNDATION

## 2.1 Literature Survey

Sentiment Analysis and Gender Prediction from Twitter data has many research consists of different tecniques and methods for evaluating and getting good rate of successful result. In this thesis, many literature surveys are examined to determine which method is suitable for the high rate success. As a result of the literature searches, it is seen that the most proper method belongs to Akshi Kumar and Teeja Mary Sebastian. In their method they put forward a model for searching and finding the sentiment through Twitter. It is a microblogging service where users post writings about any topic. They interpret a hybrid approach by dictionary and corpus based methods with the aim of specifying the semantic orientation of the tweets. In order to reveal the usage and efficiency, a case study is conducted. [2]

The challenges that arise from Twitter data streams are argued by Eibe Frank and Albert Bifet. The pioneer of these challenges are the classification problems. Afterwards for sentiment analysis and opinion mining, Twitter data streams are evaluated. In order to handle the streaming unbalanced classes, a sliding window Kappa statistic for evaluation is suggested in time-changing data streams. For data streams, they worked on Twitter data via learning algorithms based on the specified statistic. [3]

Furthermore, term distribution involved in supervised term weighting methods is preferred rather than traditional methods. This method is crucial for representing the text under text classification. The traditional and supervised methods are contrasted within different aspects for two datasets composed of Turkish Twitter posts upon the study of Mahmut Çetin and M. Fatih Amasyalı at Yıldız Teknik University. The studies show that supervised term weighting methods are more preferable and suitable to apply. [4].

Moreover a new approach which is adding semantics into training set as additional properties for sentiment analysis, is presented by Yulan He, Harith Alani and Hassan Saif. A tweet's sentiment concept is added as an extra property whereas concept correlation is evaluated via negative/positive sentiment for each accessed entity from tweets. Another comparison is with an approach founded on topic analysis relevant to sentiments. As a result, it appeared that more efficient Recall and F score is produced during negative sentiment classification as well as more efficient Precision and Recall score during negative and positive sentiment classification. [5]

Alexander Pak and Patrick Paroubek's study is on linguistic analysis of the collected corpus besides explored phenomena explanation. Their study aims to indicate how to gather a corpus automatically toward sentiment analysis and opinion mining. Corpus usage provides a sentiment classifier build which is capable of classifying the sentiments as positive, negative and neutral. According to the experiments, the technique they suggested seems to be more efficient and have better performance than the methods used prior. Although they maintained their study in English, it can be applicable to any other language [6]. Based upon the upcoming work accommodation, in order to improve the method for sentiment analysis assessment on Twitter data, this thesis is generated.

Another study on text classification belongs to Alec Go, Lei Huang and Richa Bhayani. They aim to generate an algorithm classifying Twitter messages depending on query terms. Via machine learning techniques and distant supervision, the method they come up with can attain sentiment classification to be accurate [7, 8].

A different work from researchers Yücel Sargın, Gizem Gezici and Dilek Tapucu, proposing and evaluating new features to sentiment classification, which is used in a word polarity based approach. Before estimating the overall review polarity, they analyze sentences in first place. After that they consider length of the sentences, position within the opinionated text, purity, subjectivity, irrealis content. Their analysis can be used to find sentences, conduct information about overall review polarity in a better way. Results shows, using these proposed features points a small improvement accuracy of sentiment classification. While improving classification accuracy, these sentence level features has gains in another task which is review summarization [9].

Target dependent classification based on if the sentiments are composed of positive, negative or neutral sentiments for Twitter is studied by Tiejun Zhao, Mo Yu and Ming Zhou. The state-of-the-art algorithm provides a solution for the problem, which is a query taking role as a sentiment target, by appropriating the target independent method assigning inconsequential sentiments to the specified target. In addition, the content of sentiment to be classified is ignored during the classification, as the state-of-the-art algorithms regards only the tweet to be classified. Nevertheless regarding the instant tweet is sufficient to classify sentiment as tweets are short and unclear. They claim to develop target dependent Twitter sentiment classification by some methods. These are incorporating target-dependent features and taking concerned tweets into thought. On the basis of the research results, this algorithm provides target dependent sentiment classification performance to be more advanced [10].

Furthermore, hashtag level sentiment classification and transmission of ontology based methods for analyzing sentiment in Twitter posts with better performance are worked on. Producing entire sentiment polarity automatically for a specified hashtag is experienced. This differs from sentiment analysis for traditional sentence level and document level significantly. As a result of their research, sentiment polarity of tweet with hashtag, hashtags co-occurrence relationship and verbal expression of the hashtags are the information beneficial for handling the assignment [11, 12].

Moreover another feature studied on and involved into the thesis is gender determination for Twitter data. One of these studies belongs to Na Cheng, R. Chandramouli and K.P. Subbalakshmi. They prefer 3 different machine learning tecniques which are support vector machine, Bayesian logistic regression and AdaBoost Decision Tree. The aim of these methods are determining gender according for short, gender mixed, free content based texts. Other identifiers for gender detection are function words, structural and word based features according to research results [13].

In addition, gender, verbal manner and social network relation is worked on by David Bamman, Jacob Eisenstein and Tyler Schnoebelen. In their study they used a novel corpus of 14,000 Twitter users. The relationship between gender and language affected manners and interests are investigated by aggregating Twitter feeds. Later on, as some of the individuals manners match with the statics about relationship between language and gender, and some of them contrast with, they search the ones manners of which match more with the other gender. The users whose manner match with the other gender are found to have more users from the other gender in their social networks. In this manner, same gender language markers relate with homogeneity. Matching up the quantitative methods and social methods bring out how genders are revealed as users related to audiences, topics, and mainstream gender norms [14].

In conclusion, another theory about deducing the gender detection for the individuals in Twitter from their tweet content, belongs to Clay Fink, Jonathon Kopecky and Maksym Morawskib. The Twitter users in the West African nation of Nigeria implemented supervised machine learning via attributes gained from user tweet content for classification. By only using unigram model 80% accuracy is provided in gender recognition, thus using only this model is a successful way for gender detection. When most dominant features are analyzed, it is seen that there are some distinctions between man and women both emotionally and topically. They claim that in case of the manners of Twitter users are not reliable or available, this method manifests clearer user information who has social media [15].

## 2.2 Application

New trend among researchers, using sentiment analysis on Twitter, for recognize scientific tests and its applications. The challenges unique to this problem area are mainly linked to the dominantly informal tone of the micro blogging. Pak and Paroubek used microblogs and especially Twitter as a body of the experimental data for sentiment analysis. They indicated; social media platforms are necessary and important source of emotions of the people's opinions; every people can express their ideas about related topics, text posts (tweets) grows every day with huge numbers, so collected corpus can be randomly large, besides regular users, politicians, celebrities, even country presidents' uses Twitter; Twitter owns very different user profiles, text posts can be collected form very different social groups and also interests groups. Twitter's users are from many different countries in the world.

Previous studies about sentiment analysis were usually in English. Nowadays number of users who use social networks in Turkey is increasing and sentiment analysis in Turkish has become need. Therefore, we created a dataset which includes approximately 85.000 words in inflected form and nominative. There are also 7650 special name which include both women and men name in Turkish except unisex name. The names are taken from TDK website. Detailed data set informations will be given at Chapter 3. According to these datasets, create an application which use Twitter API when it collects data from Twitter according to selected keyword.

Investigations are made in line with word(s) determined on a certain subject or person on Twitter, and Trend Analysis studies are made about that subject or person through obtained Tweets. Within context of this thesis, studies have been conducted about Galatasaray – Fenerbahçe derby match in Turkey that was performed via Twitter on 08.08.2014. On this study, 'Galatasaray' and 'Fenerbahçe' keywords have been used. Total of approximately 1.500.000 Tweets have been obtained during the day.

As mentioned above, a Tweet collection application in C# language has been developed for collecting tweets sent in Twitter according to requested keyword. Twitter API that was shared by twitter for developers for developing this application has been used. With help of this API, search studies have been made instantly according to word, and each Tweet has been catched by users with this application. Also, all these obtained Tweets have been stored by using MS SQL Database. So, all these Tweets sent by persons have been stored, and loss of them has been prevented with this application developed even they have been erased from user accounts.

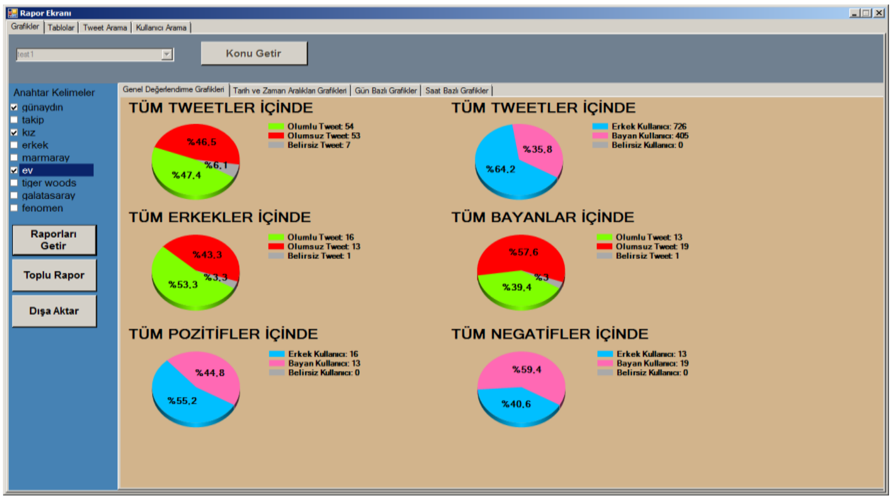


**Figure 1 General Evaluation Graphics**

After searching procedures made according to determined keyword and storing obtained data, analysis of them should be presented to user in an understandable language. This context is called as "Trend Analysis" in literature. Data required for this analysis are presented to users again via Twitter API. With help of this application developed with API, many information such as User Name, Author Name, Tweet Sending Date of user sending the Tweet, latitude and longitude information related to province of user if confidentiality settings are open for sharing, profile photo used and tweet sent are accessed. With help of that information, detailed reporting can be made about search made by user. We can arrange these reports in 4 main titles. These are; General Evaluation Graphics, Date- and Time-Interval Evaluation Graphics, Day-based Evaluation Graphics and Time-Based Evaluation Graphics. Screen images related to them are as follows.

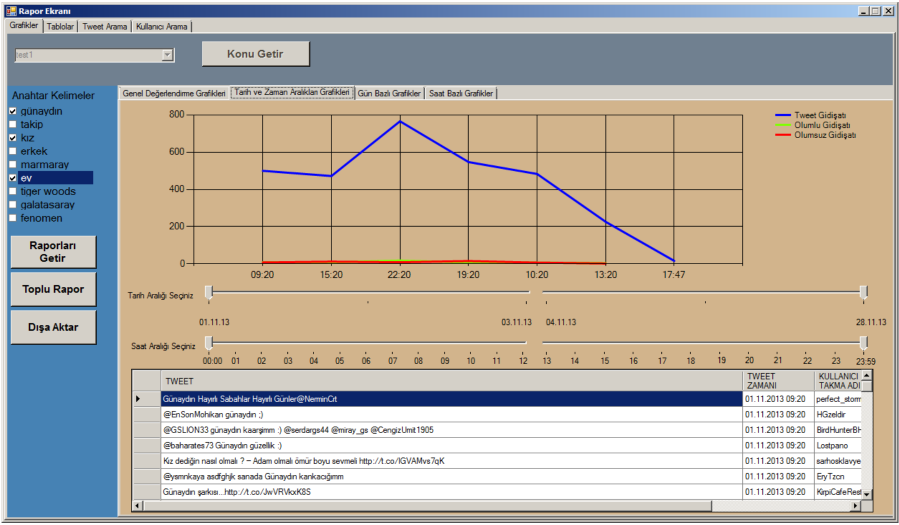


Figure 2 Date and Time Interval Evaluation Graphics

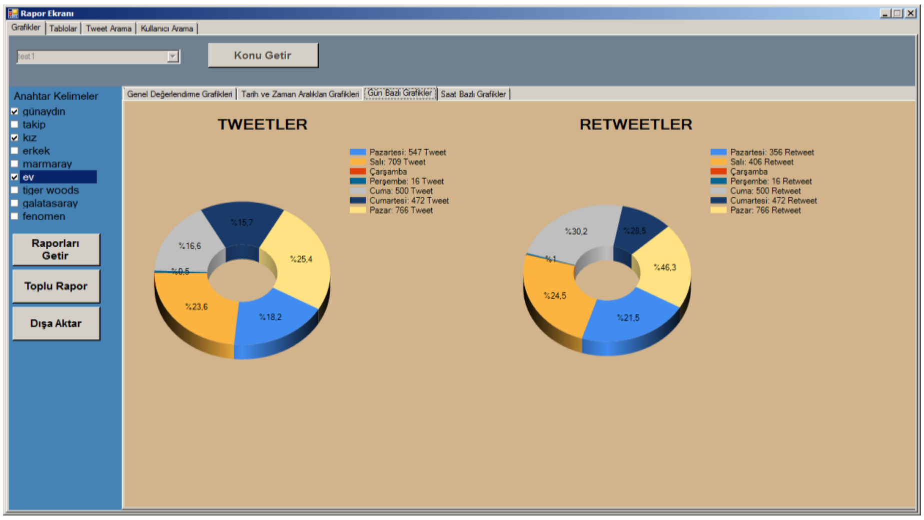


Figure 3 Day-based Evaluation Graphics

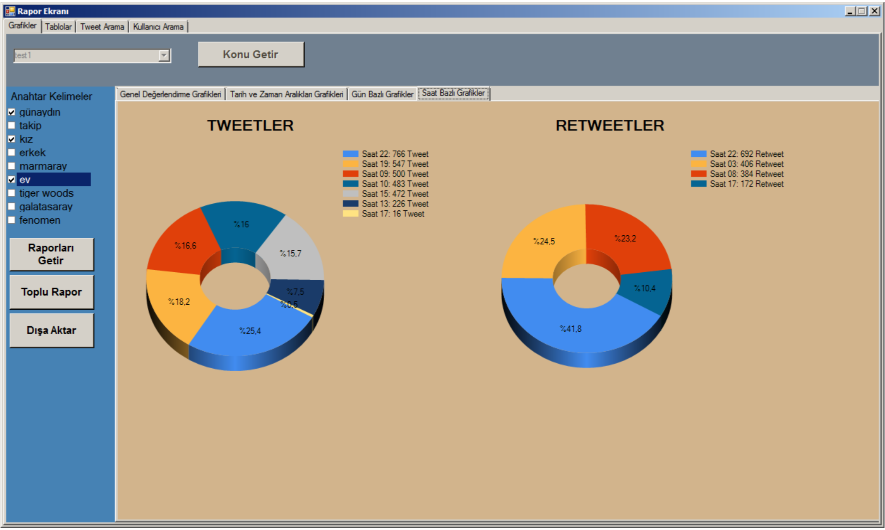


**Figure 4 Time-Based Evaluation Graphics**

Moreover, information about from which province and country users sent Tweet who doesn't hide and share their location information is obtained. For this, application developed by using Twitter API accounts latitude and longitude information shared by user sending Tweet. In addition to developed application, latitude and longitude information and provinces and countries from where Tweets were sent are determined by using developed 'GeoNames API'. So, tweets in the light of geographic location information shared by persons are reported as intensity maps on basis of provinces. A sample intensity map made on basis of Provinces of Turkey has been given on figure below.

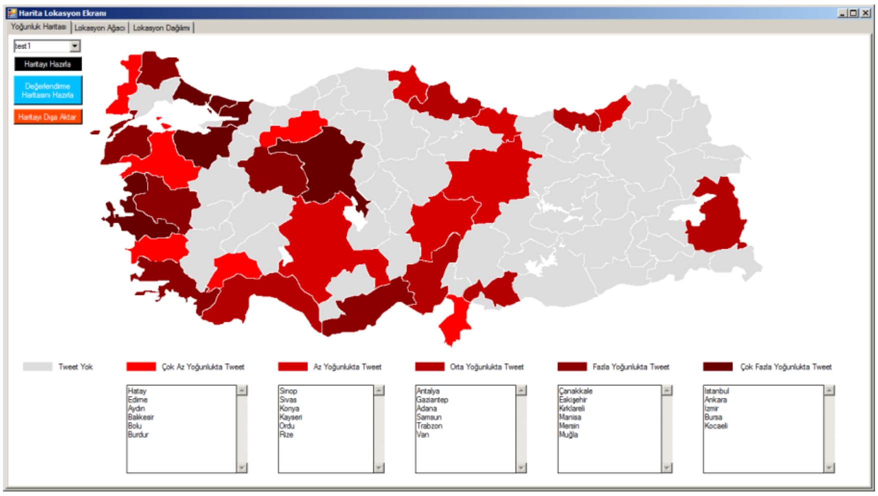


**Figure 5 Tweet Intensity on Turkey Map**

In the light of all these information, main purpose of this thesis is to perform interpretation of tweets sent as “Positive”, “Negative” or “Neutral” automatically and faster way by a computer according to determined keyword in addition to this type of information to be obtained. This subject is taken as “Twitter Sentiment Analysis” in the literature. Also, another purpose of the thesis is to perform Gender Determinations of persons automatically and faster way by taking User and Author names shared by users sending tweets. This subject is mentioned as “Twitter User Gender Determination” in the literature.

# CHAPTER 3

# METHODOLOGY

## 3.1 Data Sets

Firstly, a database containing words used by persons while expressing their emotions has been created for Sentiment Analysis studies. Studies performed in the article being examined have been executed in English Language, and it has been studied with English words and word groups. Because a Turkish Sentiment Analysis will be made in the study as a content of thesis, 117.659 words have been reached via internet belonging to SentiWordNet.

The scores of the sentiments which are both real valued and neither positive nor negative, are attached to WordNet synsets (Fellbaum1998) by SentiWordNet (Baccianella, Esuli, and Sebastiani 2010). Toward commercial applications, unrestrained distribution and licensing of non-commercial usage processes exist.

## 3.2 Measures

### 3.2.1 WordNet Dissemination

In this chapter of the thesis, Wordnet Dissemination which is the selected by human manually selected seed sets will be preserved. Table 1 below shows the words and their scores according to emotions. Negative Emotions has NegScore and Positive emotions has PosScore [16].

| Id | PosScore | NegScore | Words |
|---|---|---|---|
| 00001741 | 0.125 | 0 | Able#1 |
| 00002927 | 0 | 0 | Adaxial#1, Ventral#1 |
| 00002012 | 0 | 0 | Abaxial#1, Dorsal#2 |
| 00003566 | 0 | 0 | Emergent#2, Emerging#2 |
| 00003398 | 0 | 0 | Nascent#1 |
| 00002023 | 0 | 0.75 | Unable#1 |

**Table 1 A Fragment of the SentiWordNet Database**

The algorithm begins with *n* small, hand-crafted seed-sets and then follows WordNet relations from them, thereby expanding their size. The sets of iteration *i* are used as seed-sets for iteration *i+1*, generally after pruning any pairwise overlap between them. The algorithm is also fully shown in Figure 6.

```
WordNetSensePropagate(S, iter)
    ▷ Input
    ▷     S: a list of synsets. For example: ⟨{brilliant·s·01, n·win·01}, {sadly·r·01, gross·a·01}⟩
    ▷     iter: the number of iterations
    ▷ Output
    ▷     T: Length(S) × 1+iter synset matrix: ( S[1]          ···  iter-th propagation of S[1] )
                                             (  ⋮                          ⋮                 )
                                             ( S[Length(S)]  ···  iter-th propagation of S[n] )

initialize T: a Length(S) × 1+iter matrix such that T[i][1] = S[i] for 1 ⩽ i ⩽ 1+iter
for i ← 1 to iter
        for j ← 1 to Length(S)
                newSame ← SamePolarity(T[j][i])
                others ← ⋃_{k=1}^{Length(S)} T[k][i] for k ≠ j        ▷ The other seed-sets in this column.
                newDiff ← OtherPolarity(others)
                ▷ For the experiments, I first calculate all the propagation sets and then eliminate their
                ▷ pairwise intersection from each, to ensure no overlap.
                T[j][i + 1] ← (newSame ∪ newDiff)
return T

SamePolarity(synsets)
newsynsets ← { }
for s ∈ synsets        ▷ Synset-level relations.
        newsynsets ← newsynsets ∪ {s} ∪ AlsoSees(s) ∪ SimilarTos(s)
        for lemma ∈ Lemmas(s)        ▷ Lemma-level relations.
                for altLemma ∈ (DerivationallyRelatedForms(lemma) ∪ Pertainyms(lemma))
                        newsynsets ← newsynsets ∪ {Synset(altLemma)}
return newsynsets

OtherPolarity(synsets)
newsynsets ← { }
for s ∈ synsets
for lemma ∈ Lemmas(s)        ▷ Lemma-level relations.
        for altLemma ∈ Antonyms(lemma)
                newsynsets ← newsynsets ∪ {Synset(altLemma)}
return newsynsets
```

**Figure 6 Figure Wordnet Data Algorithm**

The algorithm uses some parameters such as the seed sets, the WordNet relationship including polarity, the iteration value, and overlap prevention. By the courtesy of these parameters the additional seed sets, extraction from other distinctions. In case of need for spreading the verbal branches, the transfer of these into the demo can be provided.

| Category | Seed Set |
|---|---|
| Female | lady, girl, female, sister, bride, women |
| Male | gentlemen, boy, male, brother, guy, man |
| Pleasure | calm, amuse, joy, ecstasy, enjoy |
| Pain | regret, agony, fearful, disconcerted, remorse |
| Strong | control, rich, perseverance, illustrious |
| Weak | weak, sluggish, lowly, poor, sorry |

**Table 2 Wordnet Category and Seed Set**

In order to polarity sense-preservation related algorithm evaluation, consequently the seed sets in Table 2 and permitted propagation algorithm will be operated. The propagation algorithm must run for 20 iterations and it controls the Positive/Negative/Neither distinctions reproduction efficiency for the subset of Harvard General Inquirer as well as in WordNet.

| **Emotions** | **Senti Words** |
|---|---|
| Negative | Bad, nasty, poor, negative, wrong, unfortunate, inferior |
| Positive | excellent, correct, positive, good, nice, superior, fortunate |
| Objective | Financial, administrative, ponder, geographic, constitute |

**Table 3 Example Senti Words**

Evaluation of the WordNet algorithm is used for the seed sets. Assessing how well simple Wordnet propagation is able to recover the Harvard Inquirer Positive/Negative/Neither classes using the seed sets of Table 3 Example Senti Words.

### 3.2.2 Weighted Wordnet Propagation

An algorithm spreading the original seed set sense as well as attaching the scores with the words and representing the intensity amount they have relevant to the graphical connections to the seed words, is suggested by Ryan, Blair-Goldensohn, McDonald and Reynar (2008).

| Seed-sets | Score vector $s_0$ | Matrix $A$ |
|---|---|---|
| $P$ (pos.)<br>$N$ (neg.)<br>$M$ (obj.) | $s_0^i = \begin{cases} +1 \text{ if } w_i \in P \\ -1 \text{ if } w_i \in N \\ \quad 0 \text{ otherwise} \end{cases}$ | $a_{i,j} = \begin{cases} 1+\lambda \text{ if } i=j \\ \quad +\lambda \text{ if } w_i \in \text{syn}(w_j) \;\&\; w_i \notin M \\ \quad -\lambda \text{ if } w_i \in \text{ant}(w_j) \;\&\; w_i \notin M \\ \qquad 0 \text{ otherwise} \end{cases}$ |

Repeated $A * s_i$: $A * s_0 = s_1$; $A * s_1 = s_2$; ... sentiment scores from the final vector (and, for each item, change its final sign to its initial sign if the two differ)

**Figure 7 Wordnet Weights Calculation**

## 3.3 Implementation

In order to have that word list become meaningful and stored in database, matching of Senti weights has been performed. Then, words having Senti Score of (Positive and Negative scores) 0 value from words in this word list have been deleted with command DELETE from SQL Table. After this procedure, 39.650 words remained consequently.

Remaining 39.650 English word and Word Groups (3N-Gram) should be translated into Turkish language in order to perform a Sentiment Analysis study in Turkish language. For this reason, Google Translate infrastructure has been used for this procedure, and their Turkish corresponding have been created with written translater application for 39.650 words. In the Turkish translation phase, Google Translate API offers you lots of same meaning words. In that words offer, 3 of the same meaning words are selected. For example “good” word has many same Turkish meanings and “iyi”, “güzel” and “hayırlı” are selected for adding our database. Words that cannot be translated have been marked with “-“. According to these information, total 30.883 words have been translated and 8767 words have been marked as “-“, e.g. they cannot be translated. Also, same words and word groups have been determined in word groups after translation, and they were excluded from database.

As mentioned above, one word has three different and same meaning Turkish words. These words are added to our database. But, it causes lots of same (not unique) words on the database. Therefore, lack of the same recording and deleting the same words and phrases in the requirement that all of the unique word has been made. As a result, 23.223 raw Turkish words and word groups have remained on our hand.

As it is known, because Turkish language is an articulate (additive) language, verbs having additions of “–mek”, “-mak”, “-me” and “-ma” in about 23 thousand words have been determined, and 48 inflection suffix of those verbs have been applied on all verbs, as in the example below for word “alkışlamak” on the sample table below.

| | PRESENT | | SIMPLE | |
|---|---|---|---|---|
| | + | - | + | - |
| **BEN** | Alkışlıyorum | Alkışlamıyorum | Alkışlarım | Alkışlamam |
| **SEN** | Alkışlıyorsun | Alkışlamıyorsun | Alkışlarsın | Alkışlamazsın |
| **O** | Alkışlıyor | Alkışlamıyor | Alkışlar | Alkışlamaz |
| **BİZ** | Alkışlıyoruz | Alkışlamıyoruz | Alkışlarız | Alkışlamayız |
| **SİZ** | Alkışlıyorsunuz | Alkışlamıyorsunuz | Alkışlarsınız | Alkışlamazsınız |
| **ONLAR** | Alkışlıyorlar | Alkışlamıyorlar | Alkışlarlar | Alkışlamazlar |

**Table 4 Example Word Conjuction_1**

| PAST | | FUTURE | |
|---|---|---|---|
| + | - | + | - |
| Alkışladım | Alkışlamadım | Alkışlayacağım | Alkışlamayacağım |
| Alkışladın | Alkışlamadın | Alkışlayacaksın | Alkışlamayacaksın |
| Alkışladı | Alkışlamadı | Alkışlayacak | Alkışlamayacak |
| Alkışladık | Alkışlamadık | Alkışlayacağız | Alkışlamayacağız |
| Alkışladınız | Alkışlamadınız | Alkışlayacaksınız | Alkışlamayacaksınız |
| Alkışladılar | Alkışlamadılar | Alkışlayacaklar | Alkışlamayacaklar |

**Table 5 Example Word Conjuction_2**

“Verbix 9 Verb Conjuctions” application has been used for this study. Total 85.313 Turkish words and word groups in database have been saved for using in Sentiment Analysis study with this application. Above example word conjuction table, word conjuctions are done for simple, present, past and future tense for both positive and negative side. With prepared Sentiment Database, sum of weights obtained by word-based matching has reached to a level to make a comment about tweets sent by persons in sentiment analysis study. You can reach picture of some part of sentiment words, word groups and their weights located in the following example.

| Id | Eng_Word | Tr_Word | Pos_Score | Neg_Score |
|---|---|---|---|---|
| 412 | above | yukarıdaki | 0 | 0.125 |
| 413 | above named | adında yukarıda | 0 | 0.125 |
| 414 | abrasive | aşındırıcı | 0 | 0.666 |
| 415 | abreact | rahatlıyorum | 0 | 0.875 |
| 416 | abreact | rahatlıyorsun | 0 | 0.875 |
| 417 | abreact | rahatlıyor | 0 | 0.875 |
| 418 | abreact | rahatlıyoruz | 0 | 0.875 |
| 419 | abreact | rahatlıyorsunuz | 0 | 0.875 |
| 420 | abreact | rahatlıyorlar | 0 | 0.875 |
| 421 | abreact | rahatlamıyorum | 0.875 | 0 |
| 422 | abreact | rahatlamıyorsun | 0.875 | 0 |
| 423 | abreact | rahatlamıyor | 0.875 | 0 |
| 424 | abreact | rahatlamıyoruz | 0.875 | 0 |
| 425 | abreact | rahatlamıyorsunuz | 0.875 | 0 |
| 426 | abreact | rahatlamıyorlar | 0.875 | 0 |
| 427 | abreact | rahatlarım | 0 | 0.875 |
| 428 | abreact | rahatlarsın | 0 | 0.875 |
| 429 | abreact | rahatlar | 0 | 0.875 |
| 430 | abreact | rahatlarız | 0 | 0.875 |

**Table 6 Conjuctions and Weights in Database**

As seen on the picture, status of words with single, double and triple words are presented in database created. With developed application, all status of words in Tweet sent has been taken up to triple groups and a Score occurs depending on their positive and negative weights and matches. In case undetermined words occur, Tweet is marked as “Neutral”.

Second aim of this thesis study, Gender Determination on Twitter will be examined under this heading. Related to this subject, primarily it is supposed that search results have been performed according to requested or determined word in Twitter. Then, tweets related to that subject or person is stored in database. As previously mentioned, URLs of user profile photos are included together with Username and Author name of Users in those information. Studies related to performing Gender Determinations of persons sending Tweet will be described in this section in line with this information of users.

Primarily, username, author name and profile photo information of users in Tweets obtained are stored in database. Then, these information are processed in a fast way and determination of users is provided as “Female”, “Male” or “Not determined”. Web page shared in internet site of TDK – Turkish Linguistic Organization containing Private Names in Turkey has been reached in order to perform this study. Names available there have been located in alphabetical order and categorized as “Female”, Male” and “Female-Male (Unisex)”. Sample data of names beginning with letter “A” is given on Table 7 below.

| | | | |
|---|---|---|---|
| Adil-(Erkek) | Adile-(Kız) | Adilhan-(Erkek) | Adlan-(Erkek) |
| Adlığ-(Erkek) | Adli-(Erkek) | Adnan-(Erkek) | Adni-(Erkek) |
| Adsız-(Erkek) | Adsoy-(Erkek) | Adviye-(Kız) | Afacan-(Erkek) |
| Afer-(Erkek) | Afet-(Kız) | Affan-(Erkek/Kız) | Afi-(Kız) |
| Afife-(Kız) | Afitap-(Kız) | Afiye-(Kız) | Afiyet-(Kız) |
| Afşar-(Erkek) | Afşin-(Erkek) | Agah-(Erkek) | Agil-(Erkek) |
| Ağa-(Erkek) | Ağabay-(Erkek) | Ağacan-(Erkek) | Ağahan-(Erkek) |
| Ağahatun-(Kız) | Ağakan-(Erkek) | Ağakatun-(Kız) | Ağan-(Erkek/Kız) |
| Ağanbegüm-(Kız) | Ağanbike-(Kız) | Ağanbüke-(Kız) | Ağaner-(Erkek) |
| Ağar-(Erkek) | Ağarantan-(Erkek) | Ağaverdi-(Erkek) | Ağbacı-(Kız) |

**Table 7 Name-Gender Table**

Based on data obtained from this site, all female and male names have been collected, and all have been stored in the same database without including in evaluation for names Female/Male “Unisex” names. According to that list, 2490 Female names, 5167 Male names have been obtained, and total 7657 names have been obtained.

Gender Prediction part of the thesis, matches have been made with names we have created separately for Female and Male names in database and determination of matched names have been performed with the application developed by considering Usernames and Author Names of users sending Tweet. With help of this developed program, fast and successful determination of Gender Prediction of users sending Tweet has been provided.

# CHAPTER 4

# RESULTS AND DISCUSSIONS

## 4.1 Sentiment Analysis Results

According to collected data from Twitter about the “Galatasaray” and Fenerbahçe keywords Figure below show the sentiment results which is automatically calculated by the application and human judgement results.

| System Score | System | Human1 | Human2 | Human3 | Human4 | Human5 |
|---|---|---|---|---|---|---|
| 0,875 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,625 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,625 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,875 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,25 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,625 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,875 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,125 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,125 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,25 | 1 | 1 | 1 | 1 | 1 | 1 |
| -1,625 | 2 | 2 | 2 | 2 | 2 | 2 |
| -0,125 | 2 | 0 | 0 | 2 | 0 | 2 |
| 0,25 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,375 | 2 | 1 | 1 | 1 | 1 | 1 |
| 1,125 | 1 | 2 | 0 | 0 | 0 | 0 |
| -0,5 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,75 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,75 | 1 | 1 | 1 | 1 | 1 | 1 |

**Table 8 FB Senti Score Result**

Figure 8 show the sentiment result of the “Fenerbahçe” keyword results. According to this table first Column name as System_Score shows the sentiment weighted score of each tweet. Meaning of these result is score which is higher than 0 has positive data for the “fenerbahçe” keyword. Score which is lower than 0 has negative data for the “fenerbahçe” keyword. If score is 0, it means system can not make comment or string matching is not successful, that’s why system can not do interpretation this tweet.

According to weights of this data can be represented by numbers that “0” for Neutral, “1” for Positive and “2” for negative meaning. These numbers are choosen because, while human do judgement on the Interpretation Screen, they select the string Positive, Negative or Neutral. It has special meaning for our application for coding. If human select Positive, application fill comments parts on the database as 1. If human select Negative, application fill comments parts on the database as 2 and 0 for Neutral.

In that part, Pearson Correlation method are used for understanding the correlation between human and system result.

| **FB – Sentiment Results** | **Pearson Correlation Between System and Human results** |
|---|---|
| Human 1 | 0.5251 |
| Human 2 | 0.5452 |
| Human 3 | 0.5133 |
| Human 4 | 0.4948 |
| Human 5 | 0.5966 |

**Table 9 Pearson Correlation Result for FB Sentiment Result**

It is the statistical method to show correlation between sets of data is a measure of how well they are related. It can also called as linear relationship between two set of data.

| System Score | System | Human1 | Human2 | Human3 | Human4 | Human5 |
|---|---|---|---|---|---|---|
| -0,25 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,25 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,5 | 2 | 1 | 1 | 1 | 1 | 1 |
| 1,5 | 1 | 2 | 2 | 2 | 2 | 2 |
| -0,125 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,375 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1,75 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,75 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0,25 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| 0,875 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,125 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0,5 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0,75 | 1 | 1 | 1 | 1 | 1 | 1 |
| -0,375 | 2 | 1 | 1 | 1 | 1 | 1 |

**Table 10 GS Senti Score Result**

In that part again, same method which is Pearson Correlation method is used for understanding the correlation between human and system result.

| GS – Sentiment Results | Pearson Correlation Between System and Human results |
|---|---|
| Human 1 | 0.561077 |
| Human 2 | 0.553637 |
| Human 3 | 0.564522 |
| Human 4 | 0.604044 |
| Human 5 | 0.558864 |

**Table 11 Pearson Correlation Result for GS Sentiment Result**

In the context of this definition, five human judgement results are compared with the system results and get some score. These score has own meaning which shown on the Table 12 Pearson Correlation Values Range.

| | |
|---|---|
| +1 | Perfect positive peak for linear relationship |
| + 0.50 | Moderate positive linear relationship |
| + 0.30 | Weak positive linear relationship |
| 0 | No linear relationship |
| – 0.30 | Weak negative linear relationship |
| – 0.50 | Moderate negative linear relationship |
| –1 | perfect negative peak for linear relationship |

**Table 12 Pearson Correlation Values Range**

The possible values for the Pearson Correlation results will be between -1 and 1. It will very rarely be 0, -1 or 1. Mostly it is a number between these values. The closer the value gets to zero, the greater the variation the data points are around line of best fit, Table 12 Pearson Correlation Values Range table explains the correlation coefficient values.

According to Correlation Value range between the -0.5 to 1.0 or 0.5 to 1.0 has higher correlation. The values between -0.3 to -0.5 or 0.3 to 0.5 has medium correlation relations. Moreover, it provides Moderate Positive Linear relationship between systems and human judgements for the sentiment analysis of both “Galatasaray” and “Fenerbahçe”.

There is also Recall - Precision method for comparing system result with the human judgement result person by person. It provides valuable result to compare human and system results judgement proximity.

| | | Human Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | a | b |
| | - | c | d |

**Table 13 Precision-Recall Result Formula**

Recall(+) = a / (a + c)  Recall(-) = d / (b + d)

Precision(+) = a / (a + b)  Precision(-) = c / (c + d)

Firstly, Formula applied for ‘Fenerbahçe’ keyword and results;

| | | Human1 Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 124 | 15 |
| | - | 12 | 40 |

**Table 14 FB-Human1 Precision-Recall Result**

Recall(+) = 124 /(124 + 12) = 0,9117  Recall(-) = 40 / (15 + 40) = 0,7272

Precision(+) = 124 / (124 + 15) = 0,8920  Precision(-) = 12 / (12 + 40) = 0,2307

| | | Human2 Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 124 | 9 |
| | - | 15 | 39 |

**Table 15 FB-Human2 Precision-Recall Result**

Recall(+) = 124 / (124 + 15) = 0,8920  Recall(-) = 39 / (9 + 39) = 0,8125

Precision(+) = 124 / (124 + 9) = 0,9323  Precision(-) = 15 / (15 + 39) = 0,2777

| | | Human3 Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 125 | 9 |
| | - | 10 | 42 |

**Table 16 FB-Human3 Precision-Recall Result**

Recall(+) = 125 / (125 + 10) = 0,9259 Recall(-) = 42 / (9 + 42) = 0,8235

Precision(+) = 125 / (125 + 9) = 0,9328 Precision(-) = 10/ (10 + 42) = 0,1923

| | | Human4 Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 127 | 4 |
| | - | 9 | 40 |

**Table 17 FB-Human4 Precision-Recall Result**

Recall(+) = 127 / (127 + 9) = 0,9338 Recall(-) = 40 / (4 + 40) = 0,9090

Precision(+) = 127 / (127 + 4) = 0,9694 Precision(-) = 9 / (9 + 40) = 0,1836

| | | Human5 Result | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 128 | 7 |
| | - | 11 | 44 |

**Table 18 FB-Human5 Precision-Recall Result**

Recall(+) = 128 / (128 + 11) = 0,9208 Recall(-) = 44 / (7 + 44) = 0,8627

Precision(+) = 128 / (128 + 7) = 0,9481 Precision(-) = 11 / (11 + 44) = 0,2000

According to each result for Human Judgements for 'Fenerbahçe' keyword;

| | Human 1 | Human 2 | Human 3 | Human 4 | Human 5 |
|---|---|---|---|---|---|
| **Recall(+)** | 0,9117 | 0,8920 | 0,9259 | 0,9338 | 0,9208 |
| **Recall(-)** | 0,7272 | 0,8125 | 0,8235 | 0,9090 | 0,8627 |
| **Precision(+)** | 0,8920 | 0,9323 | 0,9328 | 0,9694 | 0,9481 |
| **Precision(-)** | 0,2307 | 0,2777 | 0,1923 | 0,1836 | 0,2000 |

**Table 19 FB Human Results for Precision Recall**

Formula is also applied for 'Galatasaray' keyword and results;

| | | **Human1 Result** | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 132 | 13 |
| | - | 12 | 47 |

**Table 20 GS-Human1 Precision-Recall Result**

Recall(+) = 132 /(132 + 12) = 0,9166 Recall(-) = 47 / (13 + 47) = 0,7833

Precision(+) = 132 / (132 + 13) = 0,9103 Precision(-) = 12 / (12 + 47) = 0,2033

| | | **Human2 Result** | |
|---|---|---|---|
| | | + | - |
| **System Result** | + | 131 | 13 |
| | - | 17 | 43 |

**Table 21 GS-Human2 Precision-Recall Result**

Recall(+) = 131 / (131 + 15) = 0,8972 Recall(-) = 43 / (13 + 43) = 0,7166

Precision(+) = 131 / (131 + 9) = 0,9357 Precision(-) = 17 / (17 + 43) = 0,2833

| | | Human3 Result | |
|---|---|---|---|
| | | + | - |
| System Result | + | 140 | 12 |
| | - | 10 | 42 |

**Table 22 GS-Human3 Precision-Recall Result**

Recall(+) = 140 / (140 + 10) = 0,9333 Recall(-) = 42 / (12 + 42) = 0,7777

Precision(+) = 140 / (140 + 12) = 0,9210 Precision(-) = 10/ (10 + 42) = 0,1923

| | | Human4 Result | |
|---|---|---|---|
| | | + | - |
| System Result | + | 137 | 8 |
| | - | 19 | 40 |

**Table 23 GS-Human4 Precision-Recall Result**

Recall(+) = 137 / (137 + 19) = 0,8782 Recall(-) = 40 / (8 + 40) = 0,8333

Precision(+) = 137 / (137 + 8) = 0,9448 Precision(-) = 19 / (19 + 40) = 0,3220

| | | Human5 Result | |
|---|---|---|---|
| | | + | - |
| System Result | + | 141 | 9 |
| | - | 9 | 45 |

**Table 24 GS-Human5 Precision-Recall Result**

Recall(+) = 141 / (141 + 9) = 0,9400 Recall(-) = 45 / (9 + 45) = 0,8333

Precision(+) = 141 / (141 + 9) = 0,9400 Precision(-) = 9 / (9 + 45) = 0,1764

According to each result for Human Judgements for ‘Galatasaray’ keyword;

| | Human 1 | Human 2 | Human 3 | Human 4 | Human 5 |
|---|---|---|---|---|---|
| **Recall(+)** | 0,9166 | 0,8972 | 0,9333 | 0,8782 | 0,9400 |
| **Recall(-)** | 0,7833 | 0,7166 | 0,7777 | 0,8333 | 0,8333 |
| **Precision(+)** | 0,9103 | 0,9357 | 0,9210 | 0,9448 | 0,9400 |
| **Precision(-)** | 0,2033 | 0,2833 | 0,1923 | 0,3220 | 0,1764 |

**Table 25 GS Human Results for Precision Recall**

According to Table 19 FB Human Results and Table 25 GS Human Results for Precision Recall results shows the harmony between the human judgements. It is a moderate level relations between the human results. Most parts are approximately same range for the correct and incorrect decision.

### 4.2 Gender Prediction Results

According to collected data from Twitter about the “Galatasaray” and Fenerbahçe keywords table below shows the gender prediction system results and human judgement results.

| System Gender | Person1 | Person2 | Person3 | Person4 | Person5 |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 |

**Table 26 FB Gender Results**

Table 20 FB Gender Result shows the gender prediction of the “Fenerbahçe” keyword results. According to this table first Column name as System_Gender shows the system gender score of each tweet. Meaning of these result is score which is “0” has no gender prediction result for the “fenerbahçe” keyword. Score which is “1” has meaning of gender of person who send tweet is women. . If score is “2”, it means system detect the gender of tweet as men. These numbers are choosen because, while human do judgement on the Interpretation Screen, they select the string “Men”, “Women” or “Not Detected”. It has special meaning for our application for coding. If human select Women, application fill gender parts on the database as 1. If human select “Men”, application fill comments parts on the database as 2 and 0 for “Not Detected”.

| System Gender | Person1 | Person2 | Person3 | Person4 | Person5 |
|---|---|---|---|---|---|
| 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 2 | 2 | 2 | 2 | 2 |
| 0 | 2 | 2 | 2 | 2 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |

**Table 27 GS Gender Results**

In that part, again Pearson Correlation method are used for understanding the correlation between human and system result.

| FB – Gender result | Pearson Correlation Between System and Human Results |
|---|---|
| Human 1 | 0.678775306 |
| Human 2 | 0.675274129 |
| Human 3 | 0.707208783 |
| Human 4 | 0.662574044 |
| Human 5 | 0.713123983 |

**Table 28 Pearson Correlation Result for FB Gender Result**

| GS – Gender result | Pearson Correlation Between System and Human Results |
|---|---|
| Human 1 | 0.606717 |
| Human 2 | 0.58301 |
| Human 3 | 0.606717 |
| Human 4 | 0.588235 |
| Human 5 | 0.606491 |

**Table 29 Pearson Correlation Result for GS Sentiment Result**

In the context of this definition, five human judgement results are compared with the system results and get some score. These score has own meaning which shown on the figure below.

| | |
|---|---|
| +1 | Perfect positive peak for linear relationship |
| + 0.50 | Moderate positive linear relationship |
| + 0.30 | Weak positive linear relationship |
| 0 | No linear relationship |
| – 0.30 | Weak negative linear relationship |
| – 0.50 | Moderate negative linear relationship |
| –1 | perfect negative peak for linear relationship |

**Table 30 Pearson Correlation Values Range**

The results will be between -1 and 1. It is difficult to see but rarely see 0, -1 or 1. Get a number somewhere in between those values. The closer the value of r gets to zero, the greater the variation the data points are around the line of best fit.

High correlation: 0.5 to 1.0 or -0.5 to 1.0

Medium correlation: 0.3 to 0.5 or -0.3 to 0.5

According to results gender prediction for the “galatasaray” and “fenerbahçe” keywords correlations between system and human judgements are strong relation. It shows people decision are moderate level and robust.

| Average | Pearson Correlation Between System and Human Average Results |
|---|---|
| FB – Sentiment results | 0.625125 |
| FB – Gender result | 0.712397 |
| GS – Sentiment results | 0.58892 |
| GS – Gender result | 0.605876 |

**Table 31 Pearson Correlation Results Between System and Result**

A Pearson correlation values between the average of the human judges and the system results are obtained in table above. The results show the system result and the human results are strongly correlate with each other’s for both sentimental analyses and the Gender prediction results for both fenerbahçe and galatasaray keywords.

Moreover, human judgement which is applied by 5 person, do sentiment and gender prediction interpretations according to their thoughts independently. At this point, there is implementation GUI for interpretation that shown in Figure below.

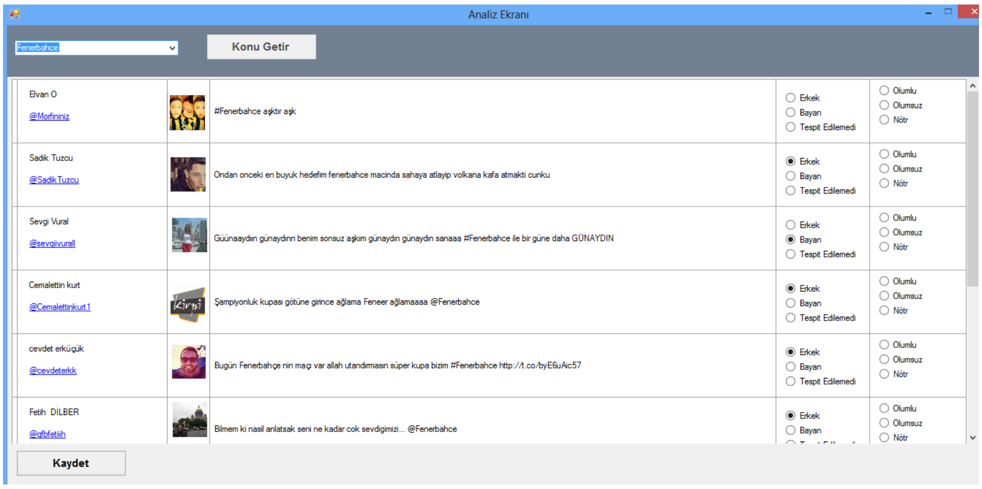


**Figure 8 Manuel Comments Screen in Application**

According to Figure, human who made a judgement about the tweet can easily apply their comments. At the sentiment analysis part, reading tweet and make comments when choose from multiple selection as “Olumlu”, “Olumsuz” and “Nötr”. At the gender prediction part. There is also helpful informations which are name, author name and profile picture of user. When human make judgement about gender prediction they see the profile picture. It is a strong evidence to understand the gender. Besides, author name and username of user also helpful for understanding the gender. Both of them shows strong relation. According to these evidences, human evaluate the tweet and make choice as “Erkek”, “Bayan” and “Tespit Edilemedi”.

It is exact that, fake accounts has bad affect on the evaluation of the gender prediction and sentiment analysis. For example, person who has fake account to not to show his or her real life and share different thoughts. Moreover, at the gender prediction part, some users name are unisex. That’s why we can not understand easily the person women or men. At that time, just looking the profile picture, sometimes it has real picture or some Twitter default picture “egg” or some other nature or landscape picture etc…

### 4.3 Fleiss' Kappa Results

At the last part of the thesis which is results and discussions, Fleiss' Kappa are applied to show the comparative reliability of the agreement between the interpreters. This method calculates the degree of agreement in classification over that which will be expected by oppurtinity. Below table shows the results about the 'Fenerbahçe' and 'Galatasaray' keywords related with the each sentiment analysis and gender prediction results.

| | GS-Sentiment | FB-Sentiment | GS-Gender | FB-Gender |
|---|---|---|---|---|
| **Kappa Results** | 0,95589 | 0,825335 | 0,981529 | 0,940812 |

**Table 32 Fleiss' Kappa Results for Gender Prediction & Sentiment Analysis**

According to result value table below Table 33 shows the value range of the Interpretation results. Kappa Results of the Sentiment Analysis and Gender Prediction values for the 'Fenerbahçe' and 'Galatasaray' keywords, human comments are almost perfect agreement with the system results.

| Result Value | Interpretation |
|---|---|
| < 0 | Poor Agreement |
| 0.01 – 0.20 | Slight Agreement |
| 0.21 – 0.40 | Fair Agreement |
| 0.41 – 0.60 | Moderate Agreement |
| 0.61 – 0.80 | Substantial Agreement |
| 0.81 – 1.00 | Almost Perfect Agreement |

**Table 33 Fleiss' Kappa Interpretation Results Value**

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

## 5.1 Conclusion

Completing successfully the “Sentiment Analysis” and “Gender Prediction” on Twitter has been aimed within context of this thesis, of which detailed explanation has been made above. For this reason, Tweets about Galatasaray – Fenerbahçe derby performed on 08.08.2014 has been studied as an example, and tweets have been collected with application developed by using keywords of “Galatasaray” and “Fenerbahçe” within this context. Approximately 1.500.00 Tweets has been reached, and 205 tweets related to keyword “Galatasaray” and 205 tweets related to keyword “Fenerbahçe” have been randomly selected from these tweets.

Then, tweets have been interpreted by using prepared single, double and triple word match methods. Also, Username and Author name of users sending tweet has been inquired and interpreted with developed application in Person Name Database. As a result of these studies;

Total 205 tweets have been examined related to keyword Galatasaray and 83 have been marked as positive, 102 as negative and 20 as Neutral with developed application.

Total 204 tweets have been examined related to keyword Fenerbahçe and 139 of them have been marked as positive, 58 as negative and 7 as Neutral with developed application. For proving correctness and success ratio of this developed application, same data have been interpreted to 5 separate persons and obtained results have been compared. As a consequence, system has moderate level ability to examine twitter sentiment analysis and gender prediction.

## 5.2 Future Work

In the basis of the thesis, Twitter sentiment analysis and gender prediction are done by developed system. According to result of the system, future work should be include to extend word dictionary to catch all word as meaningful like positive or negative. For the gender prediction part, understanding the gender of the user from his or her tweet is the improvable acitivity to understand which words are most masculine or feminine.

## REFERENCES


1. **Mike Thelwall, Kevan Buckley, and Georgios Paltoglou,** "*Sentiment in Twitter Events*", Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton

2. **Akshi Kumar and Teeja Mary Sebastian (July 2012),** "*Sentiment Analysis on Twitter*", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3.

3. **Albert Bifet and Eibe Frank**, "*Sentiment Knowledge Discovery in Twitter Streaming Data*".

4. **Mahmut Çetin ve M. Fatih Amasyalı,** "*Supervised and Traditional Term Weighting Methods for Sentiment Analysis*", "*Eğiticili ve Geleneksel Terim Ağırlıklandırma Yöntemleriyle Duygu Analizi".*

5. **Saif Hassan; He, Yulan and Alani, Harith (2012),** "*Semantic Sentiment Analysis of Twitter*", The 11th International Semantic Web Conference (ISWC 2012), 11-15 November 2012, Boston, MA, USA.

6. **Alexander Pak, Patrick Paroubek,** "*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*", pp. 1320-1326.

7. **Alec Go, Lei Huang, Richa Bhayani (June 6, 2009),** "*Twitter Sentiment Analysis*", CS224N – Final Project Report.

8. **Alec Go, Lei Huang, Richa Bhayani,** "*Twitter Sentiment Classification Using Distant Supervision*".

9. **Gizem Gezici, Berrin Yanikoglu, Dilek Tapucu and Yücel Saygın,** "*New Features for Sentiment Analysis: Do Sentences Matter?*".

10. **Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao,** "*Target-dependent Twitter Sentiment Classification*", Microsoft Research Asia.

11. **Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades (2013),** "*Ontology-based Sentiment Analysis of Twitter Posts*", Expert System with Applications.

12. **Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang,** " *Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach*", Microsoft Research Asia, Beijing, China.

13. **Na Cheng, R. Chandramouli, K.P. Subbalakshmi, (April, 2011),** "*Author Gender Identification From Text*", Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA, Digital Investigation, pp. 78-88.

14. **David Bamman, Jacob Eisenstein, Tyler Schnoebelen,** "*Gender in Twitter: Styles, Stances, and Social Networks*".

15. **Clay Fink, Jonathon Kopecky and Maksym Morawski,** "*Inferring Gender from the Content of Tweets: A Region Specific Example*".

16. **Dmitry Davidov, Oren Tsur, Ari Rappoport,** "*Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*", ICNC, Institute of Computer Science, The Hebrew University.

17. **Rohit Kumar Jha [11615] and Sakaar Khurana [10627] (November 19, 2013)**, "*Sentiment Analysis in Twitter Project Report*".

18. **Sibel Doğan ve Banu Diri,** "*Türkçe Dokümanlar için N-Gram Tabanlı Yeni Bir Sınıflandırma (Ng-ind): Yazar, Tür ve Cinsiyet*".

19. **M. Fatih Amasyalı, Sümeyra Balcı, Esra Nur Varlı, Emrah Mete,** "*Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yönetemlerinin Performans Karşılaştırması*", *"A Comparison of Text Representation Methods for Turkish Text Classification"*.

20. **Bo Pang and Lillian Lee, (2008),** "*Opinion Mining and Sentiment Analysis*", Foundations and Trends in Information Retrieval, Vol. 2, 1-2 (2008) 1-135.

21. **Sibel Doğan**, **(2006),** "*TÜRKÇE DOKÜMANLAR İÇİN N-GRAM TABANLI SINIFLANDIRMA: YAZAR, TÜR ve CİNSİYET*", Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

22. **Çiğdem Aytekin, İslam Mayda**, **(December, 2013)**" *Sosyal Medyada Rekabet Analizi İçin Karşılaştırma Görevine Yönelik Fikir Madenciliği Modeli*", Journal Academic Marketing Mysticism Online (JAMMO). Vol 7. Part 27. 414-425.

23. **Efthymios Kouloumpis, Theresa Wilson, Johanna Moore,** "*Twitter Sentiment Analysis: The Good the Bad and the OMG!*".

24. **William B. Cavnar and John M. Trenkle,** "*N-Gram-Based Text Categorization*", Environmental Research Institute of Michigan.

25. **Umut Eroğul**, **(June, 2009),** "*SENTIMENT ANALYSIS IN TURKISH*", Master of Science in Computer Engineering Department, Middle East Technical University

26. **Bo Pang and Lillian Lee,** "*A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*", Department of Computer Science Cornell University.

27. **Patrick Lai,** "*Extracting Strong Sentiment Trends from Twitter*", Computer Science Department Stanford University.

28. **R. De Groot (July 23, 2012),** "Data *Mining for Tweet Sentiment Classification*", Master Thesis, Utrecht University, Faculty of Science, Department of Information and Computing Science.

29. **Siddharth Batra and Deepak Rao,** "Entity *Based Sentiment Analysis on Twitter*", Department of Computer Science, Stanford University.

30. **Kushal Dave, Steve Lawrence and David M. Pennock,** "*Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*".

31. **Hassan Saif, Yulan He and Harith Alani,** "*Semantic Sentiment Analysis of Twitter*", Knowledge Media Institute, The Open University, United Kingdom.

32. **Subhabrata Mukherjee, Pushpak Bhattacharyya,** "*Sentiment Analysis in Twitter with Lightweight Discourse Analysis*", IBM India Research Lab, Department of Computer Science and Engineering, IIT Bombay.

33. **Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau,** "*Sentiment Analysis of Twitter Data*", Dept. of Computer Science, Columbia University, New York.

34. **Shivakumar Vaithyanathan, Bo Pang and Lillian Lee, (July, 2002),** "*Thumbs up? Sentiment Classification using Machine Learning Techniques*", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

35. **Kiran Sarvabhotla, Prasad Pingali, Vesudeva Varma,** "*Sentiment Classification: A Lexical Similarity Based Approach for Extracting Subjectivity in Documents*".

36. **Akshi Kumar, Teeja Mary Sebastian, (2012),** "*Sentiment Analysis: A Perspective on its Past, Present and Future*", I.J. Intelligent Systems and Applications, 2012, 10, pp. 1-14.

37. **Ravi Parikh and Matin Movassate, (June, 2009)** "*Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques*".

38. **Sapna Negi and Mike Rosner,** "*UoM: Using Explicit Semantic Analysis for Classifying Sentiments*", Msida, MSD2080, University of MALTA.

39. **Amaç Herdağdelen, (January, 2012),** "*Twitter N-Gram Corpus with Demographic Metadata*".

40. **Gizem Gezici, Rahim Dehkharghani, Berrin Yanikoglu, Dilek Tapucu, Yucel Saygin,** "*SU-Sentilab: A Classification System for Sentiment Analysis in Twitter*", Sabancı University, İstanbul, Turkey.

41. **Amaç Herdağdelen, (November, 2010),** "*Collecting Common Sense From Text and People*", Doctor of Philosophy in Cognitive Sciences, University of Trento.

42. **John D. Burger, John Henderson, George Kim and Guido Zarrella,** "*Discriminating Gender on Twitter*", The MITRE Corporation, 202 Burlington Road, Bedford, Massachusetts, USA.

43. **William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, Wei Hu, (2012),** "*Gender Identification on Twitter Using the Modified Balanced Winnow*", Scientific Research, pp. 189-195.

44. **Zachary Miller, Brian Dickinson, Wei Hu, (2012),** "*Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features*", Department of Computer Science, Houghton College, Houghton, USA.

45. **Jalal S. Alowibdi, Ugo A. Buy and Philip Yu, (2013),** "*Language Independent Gender Classification on Twitter*", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

46. **Amaç Herdağdelen and Marco Baroni (April, 2011),** "*Stereotypical gender actions can be extracted from Web text*", CIMeC, University of Trento, Corso Bettini 31, Rovereto, Italy

47. **Sasha Blair Goldensohn, Kerry Hannan and Ryan McDonald**, "*Building a Sentiment Summarizer for Local Service Reviews"*

48. https://raw.githubusercontent.com/rs93/nytml/master/sentiwordnet/sentiword.txt

## APPENDICES A

## CURRICULUM VITAE

### PERSONAL INFORMATION

**Surname, Name:** Balaban, Ertuğrul
**Nationality:** Turkish (TC)
**Date and Place of Birth:** 28 October 1986, Ankara
**Marital Status:** Single
**Phone:** +90 532 777 04 98
**Email:** e.balaban7@gmail.com.tr

### EDUCATION

| Degree | Institution | Year of Graduation |
|---|---|---|
| M.Sc. | Çankaya Univ., Computer Engineering | 2015 |
| B.Sc. | Bilkent Univ., Computer Tech. and Information Systems | 2010 |
| High School | Ayrancı High School | 2004 |

### WORK EXPERIENCE

| Year | Place | Enrollment |
|---|---|---|
| 2012- Present | Natek Bilişim A.Ş. | Senior Software Engineer |
| 2011 December | iSoft | Software Developer |
| 2009 October | Meteksan Savunma | Assistant Engineer |

### FOREIN LANGUAGES
Advanced English, Beginner German

## PROJECTS

1. ProsMedya - ProsMedya is a social media digger, achiever, analyzer and reporter tool.
2. Reklam Karesi - Reklam Karesi is a social responsibility project that is derived from the million dollar homepage concept.

## CERTIFICATES

1. Certified Software Process Manager
2. Professional Scrum Master
3. ISTQB Certified Foundation Level Tester

## HOBBIES

Semi-Professional Basketball, Football, Swimming and Cinema