CLASSIFICATION BASED PERSONALITY ANALYSIS ON TURKISH TWEETS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKALP MAVİŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

AUGUST 2019

Approval of the thesis:

CLASSIFICATION BASED PERSONALITY ANALYSIS ON TURKISH TWEETS

submitted by **GÖKALP MAVİŞ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** ________________

Prof. Dr. İsmail Hakkı Toroslu
Supervisor, **Computer Engineering, METU** ________________

Examining Committee Members:

Prof. Dr. Pınar Karagöz
Computer Engineering, METU ________________

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU ________________

Assoc. Prof. Dr. Osman Abul
Computer Engineering, TOBB ETU ________________

Date:

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, Surname: Gökalp Mavİş

Signature :

ABSTRACT

CLASSIFICATION BASED PERSONALITY ANALYSIS ON TURKISH TWEETS

MavİŞ, Gökalp
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. İsmail Hakkı Toroslu

August 2019, 68 pages

Psychology researches suggest that some of the personality traits correlate with linguistic behavior and nowadays people most commonly present themselves to the world by using social media. In social media, users openly reveal insights into their lives and details about their personality. Although, there are many studies related to social media, only a small number of them work on personality prediction. In this project, we used not only text data but also the other statistics of users of Twitter which is one of the most commonly used social media platforms. The intent of this thesis was modeling the correlation between Twitter data and Big Five Personality Traits by using machine learning algorithms. In this thesis, collected data, methods of analyzing them and the machine learning techniques that successfully predict personality are going to be described.

Keywords: machine learning, Twitter, personality, Big Five Personality Traits

ÖZ

TWİTTER’DAN KİSİLİK ANALİZİ

MavİŞ, Gökalp
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. İsmail Hakkı Toroslu

Ağustos 2019 , 68 sayfa

Psikoloji alanındaki araştırmalar bazı karakter özelliklerinin dilin kullanımı ile ilişkilendirilebileceğini ortaya koymuştur. Günümüzde sosyal medya kişilerin hayatlarına dair detayları ve yaşayış şekillerine dair ipuçlarını dünya ile paylaştığı platform konumuna gelmiştir. Sosyal medya üzerine bir çok araştırma olsa da bunların çok az bir kısmı sosyal medyayı kullanarak kişilik analizi yapmayı amaçlamıştır. Biz araştırmamızda Twitter’dan aldığımız dil kullanımından kaynaklı özellikleri, takip edilen kullanıcı sayısı gibi diğer Twitter istatistikleriyle zenginleştirerek kullandık. Araştırmamızın amacı, Twitter’dan aldığımız bilgiler ile "Beş Büyük" kişilik özelliği (Dışadönüklük, Duygusal Kararlılık, Anlaşılabilirlik, Dürüstlük, Yeni Tecrübelere Açıklık) arasındaki ilişkiyi makina öğrenmesi algoritmaları kullanarak bulmaktır. Bu tezde, topladığımız bilgiler, onları inceleme methodlarımız ve bunları makina öğrenmesi algoritmalarıyla kullanış şeklimiz anlatılmaktadır.

Anahtar Kelimeler: makina ogrenmesi, Twitter, personality, Bes Buyuk Kisilik Modeli

This is for you, Mom.

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor Prof. Dr. Ismail Hakki Toroslu. The door to his office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it. Besides, I greatly appreciate the feedback offered by Prof. Dr. Pinar Karagoz. ˘ She significantly helped me to improve numerous technical aspects of my study.

A special thanks to my family. Words cannot express how grateful I am to my mother, father and sister for providing me an unfailing support and a continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you all.

I would also like to thank all of my fellas who supported me in writing, and incented me to strive towards my goal. There were times when I couldn't have spare time to spend with them while I was busy with writing my thesis but they were always there for me whenever I need their company.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

LIWC	Linguistic Inquiry and Word Count
NLP	Natural Language Processing
SNS	Social Network Sites
KDD	Knowledge Discovery in Databases
FFM	Five-Factor Model of Personality
WC	Word Count
HCI	Human Computer Interaction
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
ERCC	Ensemble of Random Chains Corrected
MAE	Mean Absolute Error
LDA	Latent Dirichlet Allocation
SVR	Support Vector Regression
RMSA	Root-Mean Square Error
ERCC	Ensemble of Random Chains Corrected
API	Application Programming Interface
URL	Uniform Resource Locator
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
POS	Part of Speech
IQR	Interquartile Range
SVC	Support Vector Classification

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Social media is the place where people present their opinions in various topics. These topics can be about just users' daily thoughts or opinions about a product. Users of Huawei products are also expressing themselves in these social media platforms. The main of this project is detecting personality types of Huawei users to consider user feedback in a more logical way. The information we will provide to Huawei will be used to define their strategies to reach the users. As an example strategy, if the users which give positive and negative feedback belong to different personality groups, Huawei can decide its advertisement policy diffently for these different groups. However, how our results are going to be used is not in scope of this thesis.

In this project, natural language processing, text mining, clustering and classification techniques are going to be used to make personality analysis from Twitter data.

1.2 Proposed Methods and Models

Character analysis is a widely researched topic in psychology and there are many questionnaires in this purpose. Our main goal is matching these questionnaire results with the social media profiles of the users and using this machine learning model to identify personalities just by using social media profiles.

This project has mainly two parts:

- Natural Language Processing and Text Mining: NLP methods are used to analyze written text in social media platforms. These methods gave us meaningful information about both syntactic usage of words and meanings of them.

- Classification: Machine learning models are created by using twitter data and personality questionnaire results. Then, these models are used to detect personalities just by using twitter data.

1.3 Contributions and Novelties

Our contributions are as follows:

- Using not only text data but also other twitter information such as user's way of interacting people.
- Turkish tweets are used and characteristics of Turkish language are examined.
- LIWC is used for Turkish texts.

CHAPTER 2

RELATED WORK

Personality is one of the typical and enduring topics in psychology. Personality prediction can be basically defined as identifiying personality traits of a person by using a set of data. As the data set, naturally provided content such as information accumulated in social media or data prepared specifically for this purpose such as essays on a specified topic can be used. Current personality studies often rely on data from well-controlled and specified environments [1]. However, SNS usage is getting more and more popular and the data produced by the user can provide a valuable approach to automatically determine human personality.

Social media is one of the most commonly parts of the Internet. According to Bargh's research the reason for this popularity is because internet is important on daily life [2]. Statistics prove that 1 in every 3 minutes of time spent in the Internet is spent on social media sites[3]. In 2010, 61% of adults in USA were using SNS and this number is naturally axpected to increase.[4]. Not only adults but also young people show inclination to use these sites. According to a study made on college students, male and female students use SNS equally and they spend nearly 3 hours per day on SNS [5]. In 2005, a survey on social networking websites forecasted that there are 115 million members using SNS[6].

Although SNS is getting more popular everyday, its association with personality is a still open research area to define if users express their personality in these sites or if personality judgments made by using the data based on a person's microblog are accurate. Defining these open topics will help to improve our knowledge of the relation of social media and personality.

Nowadays, most of the studies about personality analysis works on samples provided under certain conditions. In this purpose, owner of the samples are generally leaded to talk or write on specified topics such as a recent personal loss [7], their personal future goals [8][9] or their daily events [10]. Thus, it still remains unclear what could be the result of these studies in situations more like real-world. In real world samples, people's personalities can be defined not only by using their writings on a specific topic or what they say about them, but also free form writing samples created by them [11]. Naturalistic approach is proved to be more powerful by Mehl and colleagues[12]. In this research, they have recorded speaking samples of participants' natural language usage and behavior. This researched provided good results about the relation between natural language use and personality is identified and many of them was not documented in previous researches on personality analysis in laboratory studies [12].

SNS are the topic of most of the researches on on social media use [13]. SNS can be a great venue for expressing alternate selves. It is identified that people, especially those high in social anxiety, show hidden self-aspects which they don't normally express in their everyday life in these sites[14]. People don't always show their natural selves in SNS, they also present their possible and ideal selves in SNS [15]. This can seem like a contradiction findings about accuracy of social profiles. However, this togetherness of both possible and actual selves are good to be used together and both of them can give unique clues about personality. Moreover, there are researches which prove that automated personality judgments depending on digital footprint are more correct than foodprints of the users' relatives [16] [17].

Identifying users' personality can be used for commercial purposes, social psychology or recommendation systems to build more personalized architectures to provide better user experiences.

2.1 By Amount of Social Media Use

The amount of time spent on SNS gives unique clues about users' personality according to many researches.

A study on college students stated that people who spend more time on SNS are more unlikely to feel less satisfied with their lives. This finding is interpreted as power of SNS to overcome low self-esteem and low satisfaction[13].

SNS use also correlated with Big 5 personality traits. It is indicated that three personality dimensions are relavent to use of social media: neuroticism, extraversion and openness [18].

Conscientiousness doesn't show guaranteed correlations with SNS usage because of contradictions in different researches. In a research, it is shown people high in conscientiousness are using SNS more easily[19]. However, according to Wilson's and Ryan's researches conscientiousness has opposite clues about ease of SNS use[20] [21].

Openness is generally found to be correlated with higher SNS use [22]. This correlation is also proven by another study of college students. This study states that this correlation exist because people spend more time on SNS are generally more open to new experiences[23].

Extraversion is related with many aspects of internet. Amichai states that people who is defined as extravert are not more likely to be heavy Internet users [24] compared to those individuals who are more introverted . It is also proved that extraverts are less likely to update their profile texts, post alone photographs, [25] or post on their walls [26]. However, these doesn't show that extravert people don't likely to use SNS. There are researches state that extravert people are regular SNS users especially for messaging [23] [18].

Neuroticism decrease the use of the internet but increases the tent to use SNS. Less neurotic people don't use Internet as heavy as their more neurotic counterparts[24]. Neurotic individuals are found to be more likely to use SNS [27] [26].

2.2 By Interaction with Others

SNS is a platform where users can socialize with the others. Some of these interactions such as liking someone's post or leaving a text message in someone's wall can be detected in social networking sites.

One of the interactions which can be detected is number of friends of a user. This parameter is correlated with multiple traits of the five personality dimensions. A research which used the myPersonality dataset of Facebook to detect the correlation between user's activity on Facebook and user's personality found that number of friends and neuroticism are negatively correlated[28]. Also, Schrammel detected that more opem people have more friends in social networking sites [29]. Number of friends is also found to be correlated with extraversion according to multiple researches. People who are extravert have more friends than people who can be called as introvert [30] [31]. These people can be called as popular users in also in Twitter [32].Extravert people not only have high number of friends but also they have higher quality friendships in SNS [33]. The reason for this is stated as higher self-esteem in a research[18]. People also has preferences while choosing friends in SNS. Work in [34] showed that agreeable people are chosen as friends more often because they are referred as easy to communicate.

SNS users can interact with others by using their public time lines or direct communication. Direct communication interaction is generally studied as text based feature in researches. Posting something to another user's profile is found to be correlated with just openness dimension of big five personality traits. Two researches on Facebook has proven that people who are more open to new experiences are more likely to post on others users' walls [23] [26].

Another way of interaction with other users in SNS is re-sharing the things which is shared by other user. In Twitter environment this action is called retweeting. A study was found that the users' tendency to retweet and The Big Five personality traits are correlated[35]. It proved that there are strong correlations between this frequency and user's personality.

2.3 By Text Based Features

As it is stated previously, the Internet and especially SNS usage is getting more and more popular. According to a research, primary purpose of people to go online is communication [36]. This causes huge accumulation in written text data because text message is the main way of online communication. There are many researches work on text data collected from SNS. [37] showed that personality prediction which is done automatedly by using the language in social media are even more accurate than personality judgments of users' relatives. Linguistic signs in SNS are proved to be enough to recognize individuals' personality [38]. In one of the most reputed researches in personality analysis area, linguistic features were half of the total features and the highest number of correlations has been found about linguistic cues [39].

Finding correlations of personality traits with text-based features has been studied in many researches. In [40], it is proven that free-form profile attributes give more accurate clues abour personality prediction. Thus, text based features are studied more commonly than non-text based features in the literature.

To analyze personality by using text data, there are many ways to get meaningful cues from the text. Polarity of the words or sentences or word counts can be used in any SNS platform like Facebook, Twittter or in combination of different SNS platforms. Although the majority of the studies are done in English, there are also many researches working on other languages too.

2.3.1 Non-English Texts

In literature, most of the researches about automated personality detection worked on English texts because it is easier to collect bigger amount of data. Another reason for this is that most of the tools like LIWC are built in English an they are ready to use in English without any data preparation. However, there are still many studies working on other languages in SNS.

Chineese is one of the most commonly studied languages about automated personality detection because of the amount of data can be collected. In [41], they proposed an idea to detect personality automatically by using SNS contents in Chinese language. In this purpose, they extracted features from 1766 Sina micro blog users. To create features, they configured LIWC to work with a simple Chinese dictionary. By using this configured version of LIWC and Sogou Cell Lexicons they trained their machine learning system. They found valid correlations between linguistic cues and music tastes of users. Another research which uses Chineese language and LIWC in Sina platform was also mainly focused on text features as they are 71 of their total 78 features and they found valid correlations[42]. As classfier Naive Bayes and Logistic Regression aare used in this study. In a study on Chineese Language worked on 222 Taiwan Facebook user by using an open-vocabulary approach [43]. In this research, different methods were used to compare the performances. They used term frequencies, tokenization and feature selection algorithm by using recursive feature elimination techniques. As the result, their precision was 60% by using a dataset in Chinese.

One of the studies on Twitter was working in multiple languages like English, Spanish, Dutch, and Italian tweets [44]. They tokenized the terms from the language used in Twitter and these tokenized terms are matched with enhanced version of LIWC. There are many meaningful results in this study but the best results achieved are in predicting the Openness dimension.

Another study done for Twitter texts was in Indonesian language [45]. They translated the MyPersonality dataset contents to Indonesian Language. Then, they extracted the most frequent 750 words similar to other open-vocabulary approaches. The classification algorithms used were Naive Bayes, KNN, and SVM. Higher accuracy obtained was 72.29% and it is stated that this could be increase if a native Indonesian language dataset was used.

2.3.2 English Texts

2.3.2.1 Twitter Based

Twitter is one of the mostly researched environments to detect personality automated. In previous sections, many researches has been mentioned about personality analysis from Twitter by using interactions with others and amount of Twitter use. There are also many others which uses linguistic cues from Twitter. To detect personality by using text data of twitter, there are multiple approaches and majority of them utilizes LIWC while some others using grouping text data called tweets.

In [46], tweets are grouped as positive or negative. Then classification algorithms applied to found personality dimensions and polarity of the tweets are correlated. The results were promising. However, in this research, neutral tweets were not classified. In a study on Twitter, tweets are grouped as : negative, positive and neutral[47]. In this research, it is presented as innovative to perform grouping the tweets instead of looking the whole user profile. This can be an example to do automated personality detection on groups created by an algorithm in a semi-supervised way. In this study, both supervised and semi-supervised learning approaches were used with a list of meta-attribute features in a Naive Bayes classifier. It is proved that the semi-supervised learning produce better results than supervised learning.

In [32], 335 Twitter profiles were analyzed and these profiles are divided into 5 groups: listeners who tends follow more users, popular people who has more followers, highly read people and two other user types which are decided by using two social media indices such as Klout and TIME. In this research, they worked on the correlation of personality dimensions and these five categories of microblog users. From this relation a correlation table has been created and personality was tried to be detected by using a regression with the M5 Rules algorithm. Personality traits of these user profiles were found succesfully and the results were promising.

In one of the highest reputed researches in automated personality detection area, they used tweets of 50 Twitter users' to train their model [39]. The aim of this research was predicting personality scores. In this purpose they have used 91 text based features from LIWC and MRC. In addition to that, they have done sentiment analysis on tweets. Then, they utilized Gaussian Process and ZeroR in the collected data and Mean Absolute Error was collected for each personality trait. The smallest MAE achieved was 0.119 for the Openness trait and the algortihm used was the ZeroR.

Most of the researches in this area was done by using Big 5 personality traits but there are some which works on other personality characteristics. In [48], they made the prediction of the Dark Triad Personality in Twitter environment by using classification. In total, they have chosen 337 features to predict personality. Some of these features were from Twitter statistic and most of the features were about frequency of predefined words for each user. SVM, Random Forest, J48 and Naive Bayes were used for 2,927 Twitter users from different countries. In this study, they found the correlations between these personality dimensions and Twitter users. The result with highest accuracy was that psychopath and machiavellian people tend to use words associated with anger included swear words.

2.3.2.2 Other Social Network Sites Based

Facebook Automated personality detection from Facebook profiles is a commonly researched area. There are many studies which uses non-text based features like Facebook popularity or user interactions with others. There also many researches which uses the texts in Facebook posts as it is going to be mentioned in this section.

There is a ready to use data set called MyPersonality which is created by David Stillwell by collecting huge amount of information from Facebook users. This data set contains information of over 6 million users which are from various backgrounds, age groups and cultures. Data collected is assumed as valid in the literature because the participation was totally based on voluntariness. There are many studies which uses this already created data set instead of trying to collect information to create grand truth to detect personality from Facebook. One of these studies which use MyPersonality corpus was using open-vocabulary approach to detect features by Latent Dirichlet Allocation algorithm [49]. In this study, LIWC is also utilized to create features and it has been showed that models created by LDA outperformed the model which is created by LIWC. Another study which also uses open vocabulary approach on MyPersonality dataset with both Latent Dirichlet Allocation and Support Vector Regression [50]. This study showed validity of LDA when the results are compared with SVR. Both of these studies were working on personality prediction and they showed strong relations between MyPersonality corpus and Big 5 personality traits.

There are also many other studies which collected their own data by surveys and various ways. In [51], the correlation between emotions in posts of Facebook users and their personality was investigated by considering characteristics like age and gender. In this research, closed-vocabulary approach was used with LIWC. There were, 100 features used in this study and 81 of them are text based features which are coming from LIWC. The other non-text based features were SNS features and time related statistics of Facebook. By using these features different classifiers were trained. K-Nearest Neighbor was the one with the highest precision and several interesting investigations were done in this study.

Like in [51], Extroversion is the personality trait which gives the most correlated results with the features extracted from Facebook. Another study about personality detection from Facebook by using text based features found that extrovert people are showing a great self-disclosure in their Facebook status updates [52]. As the reason for this, it is stated that extroversion is associated with self-monitoring [53] and self-consciousness [54]. Talking about hidden self-aspects is studied in another research too [24]. According to this study neurotic people also tend to show their hidden self-aspects like extrovert people because neurotic and extrovert people show similar behaviors like people high in social anxiety.

Neurotic people are people who are low in self-esteem [14]. This relation is researched in Facebook status updates too [55]. It is found that neurotic people use negative emotions via Facebook just like people low in self-esteem.

LIWC is a commonly used tool in researches about automated personality detection from Facebook. There is a study which validates the use of LIWC in these researches [56]. In this study, survey results and features extracted from LIWC are used as predictors and regression model is created to predict Five Factor personality of the users. The classified model was working with the 74.6% accuracy which can be considered as a successful result. This study showed the validity of LIWC usage in personality prediction from Facebook.

Blogs Blogs are also a good source of text based data because expressing the users' opinion is the written format is the main aim in this environment. In [57], blog text is proved as a promising way of predicting users personality.

In [58], blog posts of 694 users are studied to find relation by using words with the

Big Five personality traits. Both open and closed vocabulary approaches are used in this study to create classification models. In closed vocabulary approach 66 LIWC categories were used and in open vocabulary approach words are used individually by dividing the text. This study found most of the correlations about the openness dimension of Big Five personality traits. While openness is related with the 393 words, other traits were found to be correlated with just 30 words.

Another study which uses users2 leanings about choosing words improves Yarkoni's results by predicting personality dimensions by using selected features [59]. In this research, the n-gram usage frequencies are counted. Also utilization of stop word is considered by observing if stop words are used or omitted. These features are combines with the features extracted from LIWC and a model is classified by SVM. The results were similar to the previous study and most of the correlations found about openness dimension of Big Five personality traits.

Youtube Youtube is a video sharing platform and most of the videos there contains language use. When the data is converted into written format, it can be easily used to classify the personality by using text based features.

In a study which uses both text based and non-text based features different kinds of regression algorithms are used and the results are compared [60]. This study used data of 404 YouTube vloggers. In this study, 7 different type of feature was used and 81 of them were LIWC features by using text data. The rest of the features were gender, NRC features, SentiStrength scores and audio-video features. Then by using these features different kinds of regression algorithms were used and the results were compared. The base learner chosen in this study was decision tree algorithm. The study showed that lowest RMSE found was in conscientiousness dimension of Big Five personality traits with a value of 0.64. The algorithm which gave the best results found to be Multi-objective random forest multivariate regression algorithm.

Cross Media Many studies which uses a specific SNS platform to create dataset have been mentioned previously. These platforms might be Facebook, Twitter, blogs or Youtube. However, there are also some researches which combines two SNS platforms for collecting data to create a classifier model.

In [61] , they collected data from 175 Twitter and Facebook users who are job applicants to see if their personality can be detected in their social media contents. They observed badmouthing behaviors of the users as the feature. The expected personality traits to be detected by using badmouthing are extraversion, conscientiousness and agreeableness. This research showed that both agreeableness and conscientiousness are negatively associated with badmouthing through SNS. However, no clue has been found about the correlation between extraversion and badmouthing.

Another research which focuses on both Facebook and Twitter tried to detect the user preferences of those two platforms [62]. In this study, they observed the user actions about an advertisement which is posted in both Facebook and Twitter and revealed a differential relationship between user behaviors on both of these platforms. Personality differences has been showed by using user preferences for Facebook or Twitter. It has been found that different purposes are the reasons for people to use the same sites.

In [63], Youtube is added to Facebook and Twitter. They tried to predict personality by utilizing datasets from these three SNS platforms. The extracted features were features from LIWC, NRC and MRC features, SentiStrength features, SPLICE features, demographic characteristics of users and audio-video features for the YouTube dataset. NRC and audio-video features were extracted just from Youtube dataset but the other features were common for all of these three datasets. Both multivariate and univariate techniques were used in this research and the classification model is created by Support Vector Machine and decision tree. It is stated that the best results achieved were by using the decision tree algorithms. Cross-media personality prediction was performed successfully in all these three platforms. However, according to the results, it is showed that using three different datasets from different platforms didn't increase the accuracy of their classifiers.

A research about personality prediction by utilizing cross-media dataset uses Twitter and Instagram [64]. Combining features from Instagram was a good contribution in the literature. Linguistic features were extracted both of these platforms. In addition to the linguistic cues, statistics from Twitter and image features of Instagram were extracted.

CHAPTER 3

PRELIMINARIES

3.1 Personality Analysis

Predicting personality is a task of identifying a person's personality trait by a dataset. There are different approaches to detect personality based on different kinds of dataset. Knowing individuals' personality can create the opportunity to create systems to predict about preferences across environments and contexts and develop recommendation systems.

Traits are the first representation of a personality structure and they represent a neuropsychic structure. Personality traits can be defined in two levels in the theory of traits psychology. The first level in this theory is called trait and personal disposition is the second level. The difference between these two is that traits are shared among different individuals while the personal disposition is about a single person. Comparison between people can be studied by using traits. However, in the personal disposition only a single person can be studied. The traits can be detected by using frequency, variety and intensity of a certain behaviour that an individual shows. As an example, if a person is sarcastic, we can expect that person to do sarcastic comments in any environment like a classroom or a social networking platform.

Personality detection based on social media usage behavior and content to create personalized recommendation systems is one of the top topics of the overall Internet area. However, in literature, personality dimensions were generally learned by questionnaires. It is not logicallypossible to get personality scores of large amount of users for personalized recommendation systems. Previous works in the field of psychology and HCI has showed the importance of detecting users' personality traits and their preferences. Detecting those can help to build personalized and adaptive systems to ensure improved and better user experiences.

Recommender systems by using personality is a well researched area. Beside, the correlation between music taste and personality is well constituted area in literature [65] [66]. Deduction of personality traits from SNS profiles can let recommender systems to have better accuracy.

Advertising can be another great topic which can benefit from known personality. [67] demonstrated relation between consumer personality and marketing techniques. The results of this study shows that knowing personality can be very beneficial for advertising.

3.2 Big Five

In last decades, psychologists who work on personality analysis by using lexical approach, ended up viewing personality in five dimensions [68]. This approach is known as Big Five personality traits [69]. This model alleges that dimensions such as neuroticism, openness, extraversion, agreeableness and conscientiousness can comprise most of the structure of the personality traits. Big Five personality traits is known as the most widely accepted personality dimensions in psychology[70].

Extraversion (also known as surgency), is the degree for dimensions like activeness, pretentiousness, talkativeness [71]. Neuroticism which describes emotional stability gradetes people according to level of anxiety, depression and nerve [72]. Agreeableness is dimension to describe characteristics like gentleness and kindness [73]. Conscientiousness is about being respectful, organized, and trustworthy [74]. Openness is basically tells about openness to experience new things. This covers personality attributes like creativeness and introspectiveness [75]. These dimensions are accepted as most extensive way of abstraction of personality.

There are many researches which use Big Five personality traits. In some of them correlation between personality traits and personal values like aims, impetuses, needs and values is worked [76]. Big Five personality traits is also tested in many ways such as its stability[77], its geographic distribution[78] and it gives valid results in these researches and many others.

3.3 LIWC

The Linguistic Inquiry and Word Count tool, known as LIWC, is a widely used and well-known tool for analyzing text to evaluate psychological properties from language [10]. LIWC is the most commonly used text analyzing tool in studies which are investigating the relation between psychological variables and word use. It is accepted as accurate in the literature because it has been proven in many researches. LIWC's main aim is identify patterns related with personality traits. This is done by calculating frequencies of words into psychologically meaningful categories, such as social terms, pronouns and affect terms [79].

LIWC extracts 81 features from a text including features about standart counts like word count [80]. There are also features about psychological processes like the frequency of the words which is related to love or hate. Some of the features in LIWC are related to usage of verbs. In this purpose, LIWC counts the number of verbs which describes an action in the past or the verbs which are used in future tense. LIWC also consider personal concerns, for example, the number of words which refer to occupation. Most of the features extracted by LIWC are about linguistic dimensions. These features related to linguistic dimensions are coming from the frequency of word types such as swear words.

Linguistic Inquiry and Word Count features are tested in three different datasets from various SNS platforms such as Facebook, Twitter and Youtube. Six of the features were found to be significantly correlated with the Big Five personality traits dimensions across all the three datasets. Depending on the dataset type, these features showed different relations. For instance, word count showed a correlation with agreeableness dimension of Big Five personality traits in all three datasets. However, the relation is 0.02 in Facebook, 0.31 in Twitter and it is negatively related in Youtube as -0.11 [63].

LIWC is originally developed in English but it is translated in many languages such as Chinese, Arabic, Italian, Dutch, Korean, German, Portuguese, Norwegian, and Spanish. In all these languages, LIWC works in closed-vocabulary approach. Chinese dictionary can be a good example for the creation of LIWC tool in a different language. Chinese version of LIWC, called CLIWC, is created by master degree candidates. Final version of CLIWC is the combination of three different dictionaries. This combination is validated manually [81].

3.4 Text Mining

Data can be understood as a quantity of facts that can be any data in a database, but also it can be data which is hidden in a simple text file. Text is one of the most natural ways of storing information. A study mentioned that in average, 80% of a company's information is stored in text documents [82]. The aim of analysis of the text data is finding connections and hidden patterns in textual data. The found pattern's quality can be measured by their validity, comprehensibility for humans, usefulness and novelty. There are different methods to find new patterns but generalized models can be produced by the found connections. The usefulness of the extracted data can be measured by how much benefit it generates for the user.

Text mining is mentioned in [83] for the first time. According to this study, text mining is described as a process to extract significant patterns to improve the knowledge of text based data sources. It is also viewed as an easy way to get data that structured from irregular data patterns and has meaning[84]. The knowledge extracted is expected to be novel, valid, ultimately understandable and potentially useful. This is done by extracting non-trivial and interesting knowledge from unstructured text based data.

According to [85], text mining can be viewed as an extension of data mining or knowledge discovery from databases. However, text mining is a more complex task than data mining because it deals with text data which is inherently vague, fuzzy, uncertain and unstructured. In literature, text mining is assumed as a process with a series steps like information extraction, the use of data mining and statistical procedures [86]. It also uses techniques from information extraction like natural language processing, information retrieval and connects those extracted features with the methods and algorithms of statistics, KDD, machine learning and data mining. According to [83], text mining doesn't have one definition but there can be different definitions of text mining by considering related research area. For example, in a study text mining can be seen as a way of information extraction which is extraction of facts from texts but in another concept it is accepted as a way of data mining done on text data as they both try to find useful patterns from data. However, this study also states that text mining is not different from data mining when preprocessing is done successfully.

Text mining is a language dependent process while data mining is not dependent on language in general. To work language independently in text mining, developing text refining algorithms is essential. That is the reason that most of the studies works on text mining focus on English documents. Mining texts in other languages creates opportunity to access previously unused information and to offer a new host of opportunities.

3.5 Twitter

Microblogs can be seen as a new trend which provide a great communication channel for users to share information which they likely would not share by using channels like phone, email, weblogs or IM. Popularity of micro-blogging increased quite quickly because it allows building common ground, enhancing information sharing and maintaining the feeling of connectedness among community. Being an option for online social networking is another main reason for microblogging to gain popularity quickly. The reason for users to choose microblogging instead of blogging is it speed and brevity[87]. Thanks to microblogging users can focus on the things happening at that moment. According to [88], this speed and possibility to share thoughts right at that moment lets users to talk about their daily routines, report news or carry out conversations.

Using microblogging is an activity that people share brief text data about tiny things that are happening in their daily life such as what they are thinking, reading and experiencing [89]. People are using micro-blogging also to gain a level of cyberspace existence and to feel another way of connection with the world. Micro-blogging is enabled by a variety of SNS tools such as Facebook, Twitter, Instagram, Pownce, Jaiku.

Twitter is one of the most popular microblogging tools. It has seen a huge growth since it was launched in October, 2006. Statistics show that the number of users on Twitter are rising constantly and it reached up to 313 million users by June 2016. In Twitter, users share their daily life activities with family members, friends and co-workers. Twitter is a not only a microblogging service but also it is a service which let user for social networking. The difference of Twitter from other SNS platforms like Facebook is that reciprocity between followers is not required in Twitter. Users can follow other users and also they can be followed by other users. Users can share news, information and opinions with interested observers[88]. They can also seek knowledge and expertise in tweets.

The posted message in Twitter is called as "tweet". Tweets can be length of 140 characters as maximum. They can contain images, sounds or texts. A tweet can be shared by another user and this is called "retweet". In Twitter, 250 billion tweets are shared per day [47]. These tweets have rich opinions about various subjects that can be useful for psychologists, marketers and other people who are interested in the extraction of opinions, moods, views and attitudes. The language used in tweets is generally informal with special characters and slangs. That is a reason that automated analysis of tweets is more difficult than more formal texts which have a better language usage and more characters.

According to [90]'s dual-factor model, people are motivated to use Twitter by two needs: self-presentation and belonging. Self-presentation is the main reason for Twitter use. Twitter activities that accomplish self-presentational purposes include wall content, profile information, and posting photographs [91]. Users who are seeking popularity tend to engage in strategic self-presentation, disclose information and enhance their profiles on Twitter [92]. Therewithal, in [93], it is showed that Twitter profiles generally represent real self-presentation of the users. Belonging is the second reason to use Twitter because it maintains and form the relationships [94]. Twitter lets users to carry out their belonging needs through learning about others and communicating. Twitter can be a good platform to cope with feelings of social disconnection because it enables relationship development and peer acceptance [95]).

The monitoring of Twitter profiles of the users is possible thanks to API that is provided by Twitter. This creates a way to access to user information and data retrieval. This API is divided into Stream API and Search API. The Stream API is a way for accessing real time messages as it can be understood from its name. On the other hand, the Search API is the way of accessing set of recent tweets. There are several libraries available in different programming languages to access to the API provided by Twitter [96].

Twitter have created the opportunity for collection of huge amounts of linguistic data which reveal users social behaviors and characteristics [97]. This data can be used in several areas like predicting user preferences across environments and contexts and develop recommendation systems.

CHAPTER 4

METHODS

4.1 Data Collection

To create our data set, we have done two things. Firstly, we administered a 60 question version of the Big Five Personality Inventory to volunteers. In the questionnaire, we have also collected their demographic information. Then, by using Twitter usernames of the volunteers, we have collected their tweets and Twitter information.

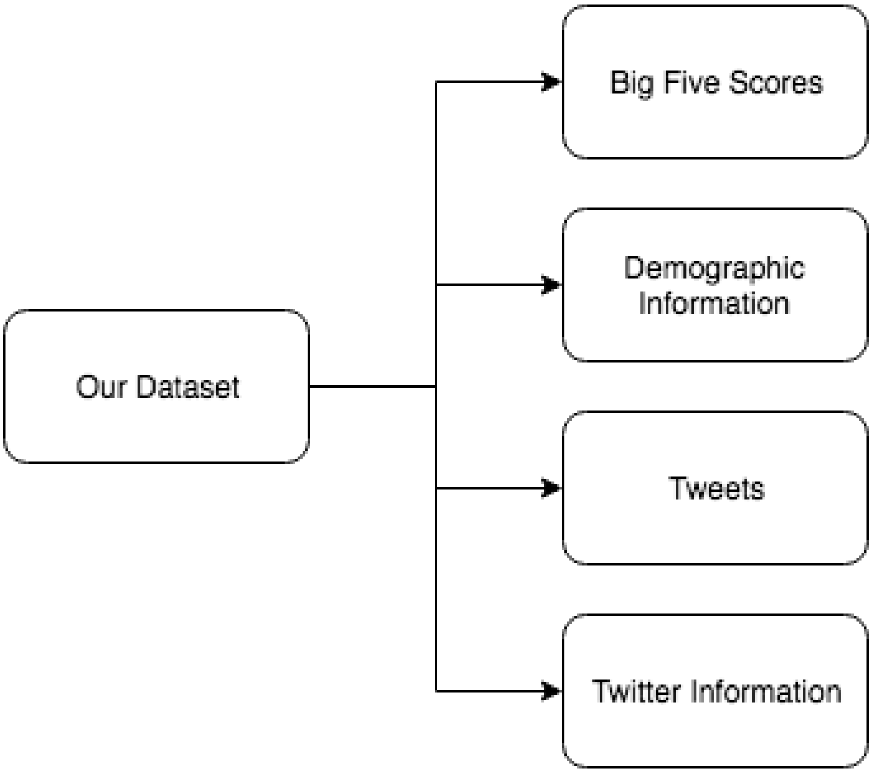


Figure 4.1: Dataset Details

4.1.1 Big Five Personality Trait Scores

Personality is traditionally measured by answering a series of questions. The questionaries which are used in this purpose generally contains from 20 to 360 questions and this is stated as the best way to measure the personality[98].

To have a valid ground truth, we needed accurate Big Five Personality traits results of the people whose tweets we are using. In this purpose, instead of creating our own questionnaire and algorithms to calculate Big Five results, we have used an already proven questionnaire which automatically calculates accurate Big Five results.

During our literature survey, we have tried to find the most accurate and scientific Big Five questionnaire. The questionnaire we have found was based on the most modern researches on Big Five Personality traits and it was designed in the professional research settings. However, it didn't provide an API to use it in a programmatic way. Also, the algorithms used to calculate the Big Five results according to this questionnaire was not provided. On the other hand, there isn't a Turkish Big Five Personality questionnaire which is proven as accurate in the literature.

To use the questionnaire we have found in our literature survey, we have developed an alternative way:

- Firstly, we needed to translate the questionnaire to Turkish because the found questionnaire was in English. The translation was done manually.
- Then, Google Forms is used as the environment to serve the translated questionnaire.
- Results were exported from Google Forms in Microsoft Excel format.
- Exported results we got from Google Forms were represented as numbers such as 1 represents the first choice in the related question of questionnaire, 2 represents the second choice, etc. By using Selenium application these results were reflected into website of the questionnaire.
- By using Selenium, we made questionnaire to produce Big Five Personality Traits by its proven algorithm. To do this, the website is manipulated by trig-

gering the right flow by Selenium.

- The results produced by the questionnaire were saved into text format by a Python script. These text files are saved by using the timestamp we got from Google Forms and this timestamp is expected to be unique.

The obtained result has scores for each dimension of Big Five Personality Traits(Neuroticism, Openness, Extraversion, Agreeableness and Conscientiousnes). These values were represented as percentage.

The questionnaire we have used contains 60 questions. In other words, there were twelve questions for each of the Big Five Personality Traits dimension in the questionnaire we have used. Each question had 5 possible choices. The first one represents that the corresponding statement is inaccurate for the user, fifth one states that user finds the corresponding question as accurate for himself/herself and the third one is the choice to select when user is neutral to the question.

The score distribution for each dimension of Big Five Personality Traits is shown in Table 4.1.

Table 4.1: Big Five Score Distribution

	0-25	25-50	50-75	75-100
Openness	0	0	15	25
Conscientiousnes	0	13	18	9
Extraversion	4	18	12	6
Agreeableness	0	9	31	0
Neuroticism	0	10	23	7

4.1.2 Demographic Information

Demographics can be described as the characteristics of a population. These characteristics can be ethnicity, race, gender, age, ethnicity, education, profession, occupation, income level.

To decide which demographic features of the participants we can use, we have checked the literature in details. Our aim was to hinder asking unnecessary things to the participants in our survey.

Many of the demographic information such as race and ethnicity was common for our participants because all of our participants were Turkish with the same ethnic background.

Some of the demographic information such as profession and occupation found to be non-affective on personality according to the studies which works on correlations between demographic information and Big Five dimensions [51].

The only demographic information we have asked in our questionnaire was age and gender.

4.1.3 Tweets

In our questionnaire, we have asked volunteers their Twitter usernames. Monitoring of Twitter profiles of the users is possible thanks to API that is provided by Twitter by using username of the user. We have used Search API of Twitter. The Search API is the way of accessing set of recent tweets of a user. We have utilized this API by Python scripts.

In Twitter, users can have two types of account either private or public. Public accounts let all users to see the whole profile including tweets and user information. However, private accounts can't be viewed by the users who are not following it. Also, the API that we used to collect user tweets were not able to get data from private accounts. Thus, we couldn't include the users with private accounts into our dataset.

The API we have used was not capable to get all tweets of the user. It allows to get tweets by the chunks of 200 tweets. After giving each chunk API also gives the id of the last tweet of the chunk and this allows to get the next chunk by using the last tweet id of the last group of tweets. This let us to get all the tweets of the user by several requests.

In total, we have collected 125000 tweets. We didn't get only text data of tweets but also;

- Text data of tweet.
- Tweet id.
- The date when the tweet is posted.
- If the tweet is retweeted or it is posted by the user.
- Link of the image shared in the tweet.
- Id of the tweet which the corresponding tweet is answer of.

In this process, we didn't get limited amount of tweets of the user. Instead of that, we have collected all tweets posted by the user. Although, this was a risk to collect unnecessary information, we wanted to train our model both by limited amount of tweets and all tweets posted by the user.

The challenge we have faced while getting the tweets was the version update of Twitter. Previous version of Twitter has a size limit for a tweet and the API was designed to get only 140 characters of a tweet. By configuring the API, we have collected text data of the tweets in any length.

4.1.4 Twitter Information

By using Twitter username we got by questionnaire, we have also collected Twitter information of the participants. We have used again the Search API of Twitter as it is explained in the previous section.

The information we have collected contains:

- Number of users followed
- Number of followers
- Number of tweets liked
- Number of tweets retweeted
- Number of total tweets
- Profile location
- Profile description
- If user has default profile picture
- If user has extended profile
- If user has a profile background image

4.2 Pre-processing

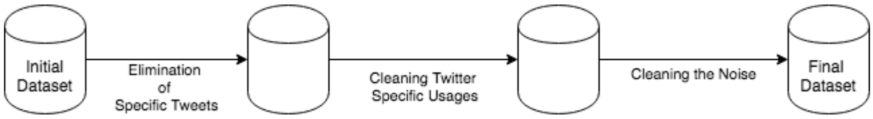

Initial Dataset
Elimination of Specific Tweets
Cleaning Twitter Specific Usages
Cleaning the Noise
Final Dataset

Figure 4.2: Preprocessing Steps

Because we have collected all tweets posted by the user, these tweets needed a pre-processing phase and some of tweets needed to be eliminated.

The tweets eliminated are:

- Tweets which are retweeted
- Tweets which are posted as an answer to another tweet
- Tweets which contains URL inside
- Tweets which mentions about another user

After this process, we had a dataset with less textual data but the cleansed dataset became more proper for text mining. NLP tools are not developed to be appropriate for the linguistic data of social media because of special usages in SNS. NLP tools are appropriate for proses but in SNS linguistic data can contain URL or it can mention about another user unlike proses.

Retweet means posting tweet of another user. Although users who shares another users tweet generally agrees with the shared tweet, sometimes it can be shared just because user finds it funny or informative. Also, even if the user completely agrees with the shared tweet, it is not said by the user's own words and the way an opinion is expressed is determiner for the tools we are using. There are multiple ways of expressing a feeling and the way it is expressed is arbiter for the tools which works with word count such as LIWC. For example, instead of the word "happy" someone can use "not sad" and the word which defines this negativity can cause different results for the NLP tools. Also typos can be used as a feature and retweets can't be used in this way because they are said by another user. Because of these reasons, we eliminated retweets from our dataset.

In Twitter, it is possible to post a tweet as an answer to another tweet. These tweets can be just a reaction by one word without using a sentence or they can be written form of laughter. Also, these answers can be used just to tag another user to make the user to see the shared tweet. By considering all these, we decided not to use the tweets which are posted as answer to another tweet.

In Twitter, it is common to share links to the photos or news. While sharing the links, users generally write their opinions about the links. However, what is written in these tweets is posted by considering that who is reading the tweet has already learned what the shared link expresses. We found these tweets useless in our model creation because it is not possible for the tools to know the content of shared link.

Tweets can mention about another user and this usage is really common in Twitter. However, in plain texts, there is no such usage. Although this is not a factor to misguide the tools which works with word count, this usage impair the integrity of meaning of the textual data. To ensure the tools which we have utilized to work properly, we didn't use the tweets which mentions about another user in our dataset.

After the extraction of some of the tweets, there were still remaining special usages in tweets. Hashtag means a phrase or a word preceded by a hash sign, widely used on social media applications and websites. Twitter is the platform where hastags are used the most commonly. The aim of using them in Twitter is identifying messages on a specific topic. Using hashtag is beneficial in Twitter because users can search for status updates with a specific hashtag, so users can categorize content and track topics on Twitter. The most common usage of hashtag is putting it into tweet at the end or beginning of tweets without attaching it to the sentences used in the tweets. Thus, deleting hashtags in the tweets doesn't change the meaning. By considering that we decided extracting the tweets which contains hashtag could cause a data loss. Instead of that, we have combed out the hashtags from the tweets and kept the cleaned tweets in the dataset.

After the data is cleaned from Twitter specific usages, there were still things to be cleaned such as multiple spaces, new lines and stop words. Firstly we cleaned multiple spaces and new lines to get a sentence which won't cause any programming errors in following processes. Then we removed stop words from the tweets because they exist in nearly all tweets and doesn't give meaningful results.

This clean data still contained noise inside such as meaningless texts. In tweets, it is quite common to have meaningless texts because of Twitter's informal usage and prevalence of slang words. This informal usage causes both grammar and spelling mistakes. To use NLP tools, we needed to normalize the tweets by correcting those mistakes because writing mistakes causes NLP tools to find wrong tags. For example, a tweet like "Bn kosmaya gidicem" has multiple spelling mistakes which can be meaningful in informal usage. In order to do this, we have used SpellChecker tool of Zemberek. Thanks to this tool, we were able to replaces the wrong words with their best alternatives. This tool gives "Ben kosmaya gideceğim" if we give the example sentence as input.

It is also common to have Turkish words written by English characters in tweets and these words causes wrong taggings while using NLP tools. To correct these mistakes we have used Zemberek's deasciification tool.

At the end of extraction and cleaning process we had 8567 tweets left which can be used as a plain text which is suitable for NLP tools and tools which works by word count such as LIWC.

4.3 Vector Construction

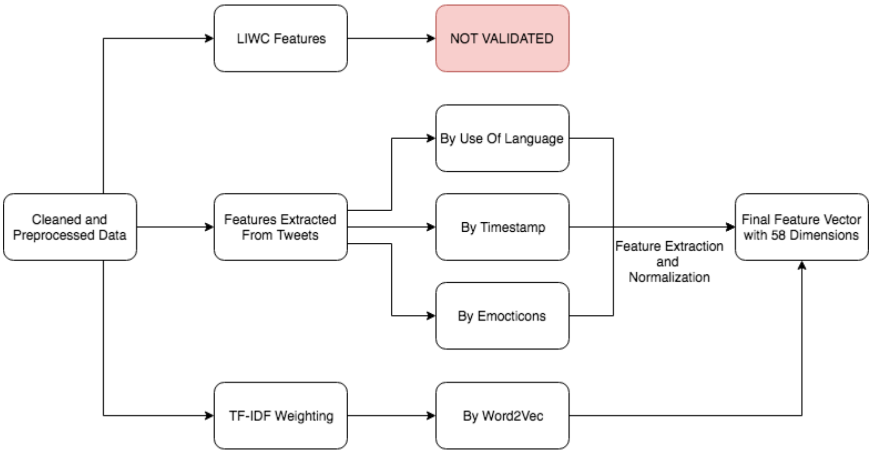

LIWC Features
NOT VALIDATED
By Use Of Language
Cleaned and Preprocessed Data
Features Extracted From Tweets
By Timestamp
Feature Extraction and Normalization
Final Feature Vector with 58 Dimensions
By Emocticons
TF-IDF Weighting
By Word2Vec

Figure 4.3: Vector Construction Steps

4.3.1 Features from LIWC

LIWC is a tool for analyzing text to evaluate psychological properties from language. It extracts 81 features from a text including features about standard counts and psychological processes like the frequency of the words.

LIWC is developed to work in English than suitable versions have been created for different languages. However, this conversion of the tool is not just translating the dictionary because LIWC also uses the semantics of the language. We have contacted the LIWC team for collaboration to translate the tool into Turkish. Estimated work in this purpose required one year work of 4 researchers including 2 linguistic scientists. This made us to find alternative ways to use LIWC.

To use LIWC, we have first translated the tweets into English by using Yandex Translate API. Than, by using python scripts, we have aggregated all translated tweets of the user into a single paragraph. Finally, we have analyzed these texts by LIWC and created features in 81 dimensions. LIWC features are listed in Table 4.2.

4.3.2 Other Features

We have collected 48 features by using use of language, terms within the texts and timestamp of tweets.

Use of Language In this manner we have collected 33 features from text including features about standard counts like word count. There are also features about statement of words like the frequency of the words which can be classified as positive or negative. We have also counted the words in different types such as adjective, verb and noun.

In order to calculate these features, we have used Part of Speech tagging by Zemberek-NLP. Then we have calculated the averages for each user to get the word frequencies to use as the weight of the feature.

The word categories we have used as feature are listed in Table 4.3.

Table 4.3: Features From use of Language

Feature	Range	Feature	Range
Case Ratio	0.58 - 1.0	Word	3.4 - 20.0
Verb	0.0 - 3.0	Noun	1.2 - 11.0
Punctuation	0.0 - 5.38	Adjective	0.0 - 2.5
Adverb	0.0 - 1.5	Numeral	0.0 - 2.06
Determiner	0.0 - 1.0	Post Positive	0.0 - 0.726
Duplicator	0.0 - 0.059	Conjunction	0.0 - 1.0
Interjection	0.0 - 1.0	Pronoun	0.0 - 1.23
Question	0.0 - 1.0	Incorrect	0.0 - 6.58
Negative	0.0 - 1.0	Plural	0.0 - 3.52
Present Time	0.0 - 1.0	Future Time	0.0 - 0.5
Past Time	0.0 - 0.88	Narrative Time	0.0 - 0.55
Progressive Time	0.0 - 1.5	Condition	0.0 - 0.5
Imperative	0.0 - 1.0	Necessity	0.0 - 0.42
Ability	0.0 - 0.5	Negative Ability	0.0 - 0.33
Question	0.0 - 1.0	Exclamation	0.0 - 0.67
Ellipsis	0.0 - 0.84	Full Stop	0.0 - 0.91
Non-Turkish Words	0.0 - 1.0		

Timestamp Timestamp of tweets gave us 4 features such as Morning, Afternoon, Evening and Night. These features represents the rate of the tweets which are posted in the specific interval. Because these four time periods are following each other in a circular manner, we wanted to specify the distances between these time periods by using two hot encoder. In this method, the distance between afternoon and evening is specified as less than the difference between evening and morning. This helped our classification algorithms to have better results.

As example:

- If the tweet is posted in Morning →Both Morning and Night columns are increased.
- If the tweet is posted in Afternoon →Both Morning and Afternoon columns are increased.
- If the tweet is posted in Evening →Both Afternoon and Evening columns are increased.
- If the tweet is posted in Night →Both Evening and Night columns are increased.

At the end we got the average of each timestamp feature.

Emoticons The tool we have used to collect tweets of users returns hexadecimal representations of emoticons. Firstly, we have decided emoticons which can be used to analyse personality. Then we have created a dictionary with the most popular emoticons by grouping related ones. We have ended up with 11 groups;

- smiling
- affection
- tongue
- neutral
- unwell
- negative
- romantic
- fingers
- activity
- sport
- plant

Each group in this dictionary represents a feature in our user vector. For each user, we have counted the emoticons used in each group and average emoticon selection of user is used as a feature.

Twitter Information From the Twitter information of users we have used.

- Number of users followed/Number of followers ratio
- Number of tweets retweeted
- Number of total tweets

4.3.3 Feature Extraction

After calculating each feature, features which are not discriminative needed to eliminated. In order to do that, we have detected features which are nearly the same for all the users.

In this purpose, we have used Scikit-learn's VarianceThreshold. Firstly, we have calculated variances for all the features and the determined threshold was 0.01. Thanks to that, we have eliminated features which have variance lower than 0.01.

After this process we have ended up with 20 features which are distinguishing. These features are; Evening, Night, Morning, Afternoon, Word, Adjective, Adverb, Noun, Verb, Plural, Full Stop, Pronoun, Incorrect, Punctuation, Numeral, Determiner, Conjunction, Negative, Negative Emoji, Smiling Emoji.

4.3.4 Normalization

In order to find the optimum one, we have tries different normalization techniques which are robust scaling, standard scaling and discretization. For robust scaling, standard scaling, we have used scikit-learn library directly. For discretization, we have implemented our own algorithim as it is going to be described in the related paragraph.

Robust Scaling Interquartile Range(IQR) is used for scaling in the robust scaling method. This method is robust against outliers. For each feature, we have calculated first and third quartiles and the median. We have calculated scaling score by removing median then we have scaled the result between first and third quartiles.

Standardization Scaling In this purpose, we have calculated means and standard deviations for each feature. Then, we have calculated z-score of each sample by subtracting the mean and dividing the result by the standard deviation. The normalized features had 1 as new standard deviation and 0 as new mean. In this method, outliers affected the resulting data more than Robust Scaling.

Discretization In the discretization, the aim is distributing continuous values into a selected number of bins to make the variable discrete-like. In our case we have chosen 4 as number of bins. To divided the data into sub-intervals, we have used predefined threshold values. Then, we have assigned new values to the intervals. Treshold values and assigned values are shown in Algorithm 1

Algorithm 1 Discretization Algorithm

```
input  : FeatureColumn
output: DiscretizatedColumn
firstQuartile = CalculateFirstQuartile(FeatureColumn);
secondQuartile = CalculateMean(FeatureColumn);
thirdQuartile = CalculateThirdQuartile(FeatureColumn);
while i=0 ; i<SizeOf(FeatureColumn) ; i++ do
    if FeatureColumn[i] <= firstQuartile then
        value = 0.0;
    else if FeatureColumn[i] > firstQuartile and FeatureColumn[i] <= secondQuartile then
        value = 0.25;
    else if FeatureColumn[i] > secondQuartile and FeatureColumn[i] <= thirdQuartile then
        value = 0.75;;
    else
        value = 1.0;
    FeatureColumn[i] = value
end
```

Performances of each normalization method is analyzed by the other project member and discretization found to give the best results. Thus, discretization is used in the following parts of our work.

4.3.5 TF-IDF Weighting

In addition to the previously explained features, we have used the terms with the tweets of users. Before applying related algorithms,we have applied two preprocessing steps; lemmatization and tokenization.

Lemmatization The main aim in lemmatization is replacing the words with their lemmas to have the same word for the different forms of it. These different forms can be because of negativity, tense or plurality of the words. We need lemmatization because words should be represented in their real dictionary forms. We have used lemmas instead of stems because paragoges used in Turkish can change the word and stemming can be not enough in these cases. For example, "Bardağım" will turn into "Bardağ" after stemming but stemming gives the result "Bardak" and this form of the word is the form that we need to use in our dictionary.

Tokenization After the lemmatization, the tweets were ready to tokenize. These tokens might be big like a sentence or a paragraph or small like a a word or a phrase. The first step of tokenization was splitting the tweets by using white spaces. To detect important phrases, we have applied another tokenization method during TF-IDF weighting with n-grams.

We have used TF-IDF vectorization to find important words or phrases in documents which we have created by appending all tweets of the user. BY crawling this document, we have calculate TF-IDF values for 1-grams, 2-grams and 3-grams. Then, by using vectorization, we have extracted top used phrases and words from the result matrix. This is a process which can also help us to learn topics the user tweeted about.

4.3.6 Word2Vec based Word Embedding

After we have chosen the words with highest TF-IDF values, we have constructed Word2Vec embedding on them by using Gensim. Gensim is a python library which is used in topic modelling. Our Word2Vec model has 38 dimensionalities. After trying different window sizes, we have detected that 7 is the optimum windows size in our case. Window size is the distance between guessed and present words in tweets. Also, we have chosen 3 as minimum window count which we have used as the threshold to ignore the words whose frequency is lower than it. After concatenation we have ended up with 38 features.

After finding Word2Vec representations of the top terms, we have concatenated them with the features found in 4.3.2. At the end, we have obtained vector with 58 features for each user.

An example row of our feature vector is;

[0.0, 0.039, -0.1502, -0.2469, -0.1361, -0.0108, 0.1292, 0.2894, -0.0672, -0.0669, 0.1206, 0.0339, -0.0018, -0.3789, -0.0888, 0.0377, -0.0128, -0.1066, 0.3264, 0.1443, 0.2707, 0.1387, 0.0156, -0.0036, 0.1216, -0.3078, 0.1636, 0.4188, 0.1409, 0.0003, 0.0118, 0.0757, 0.1052, -0.1293, 0.0922, -0.0963, 0.0356, -0.0487, -0.8425, -0.033, -0.2229, 0.033, 0.2229, 0.0389, -0.1753, 0.2311, -0.3867, -0.2606, 0.0169, -0.0126, -0.0107, -0.0118, 0.1246, 1.2662, 0.0309, -0.4488, -0.1079, -0.3878, 0.0]

Table 4.2: LIWC Features

Word count	Analytical thinking	Clout
Authentic	Emotional tone	Words/sentence
Words > 6 letters	Dictionary words	Total function words
Causation	Discrepancy	Tentative
Certainty	Differentiation	Perceptual processes
See	Hear	Feel
Biological processes	Body	Health
Sexual	Ingestion	Drives
Affiliation	Achievement	Power
Reward	Risk	Past focus
Total pronouns	Personal pronouns	1st person singular
1st person plural	2nd person	3rd person singular
3rd person plural	Impersonal pronouns	Articles
Prepositions	Auxiliary verbs	Common Adverbs
Conjunctions	Negations	Common verbs
Common adjectives	Comparisons	Interrogatives
Numbers	Quantifiers	Affective processes
Positive emotion	Negative emotion	Anxiety
Anger	Sadness	Social processes
Family	Friends	Female references
Male references	Cognitive processes	Insight
Present focus	Future focus	Relativity
Motion	Space	Time
Work	Leisure	Home
Money	Religion	Death
Informal	Swear words	Netspeak
Assent	Nonfluencies	Fillers

CHAPTER 5

EXPERIMENTS

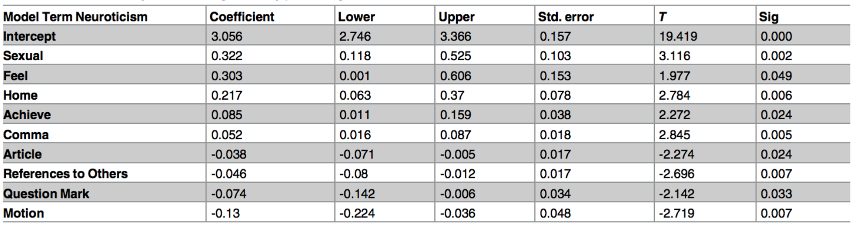
5.1 Classification

5.1.1 Using LIWC Features

Because we don't have a Turkish version of LIWC, first we needed to verify if the method we use to utilize LIWC is accurate or not. To utilize LIWC on Turkish text, we have translated the texts from Turkish to English. This translation might cause LIWC to work wrong because we didn't use the semantics of Turkish. In order to validate the LIWC features calculated by this way, we first tested our features by using already proven methods and compared the results. In this purpose, we used [56]'s results as base.

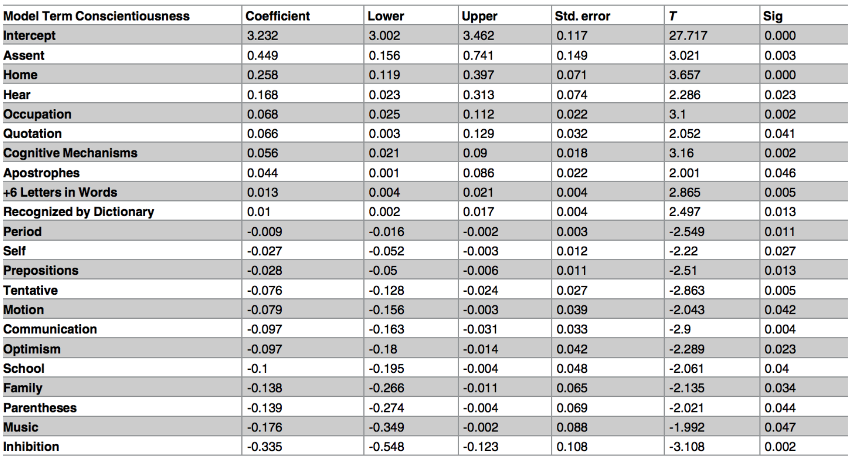
In [56], they worked on the same problem with us, personality analysis from social media texts. However, in this research, they worked on English texts by using original version of LIWC. In this study, survey results and features extracted from LIWC are used as predictors and regression model is created to predict Five Factor personality of the users. The classified model was working with the 74.6% accuracy which can be considered as accurate. This study showed the validity of LIWC usage in personality prediction from social media texts.

The models we have used are showed in Figure 5.1, Figure 5.2, Figure 5.3, Figure 5.4 and Figure 5.5.



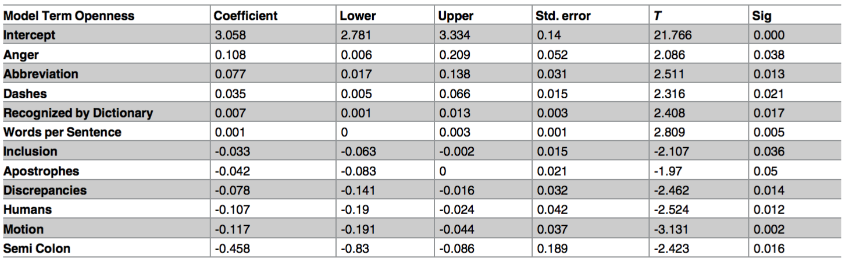
Model Term Neuroticism	Coefficient	Lower	Upper	Std. error	*T*	Sig
Intercept	3.056	2.746	3.366	0.157	19.419	0.000
Sexual	0.322	0.118	0.525	0.103	3.116	0.002
Feel	0.303	0.001	0.606	0.153	1.977	0.049
Home	0.217	0.063	0.37	0.078	2.784	0.006
Achieve	0.085	0.011	0.159	0.038	2.272	0.024
Comma	0.052	0.016	0.087	0.018	2.845	0.005
Article	-0.038	-0.071	-0.005	0.017	-2.274	0.024
References to Others	-0.046	-0.08	-0.012	0.017	-2.696	0.007
Question Mark	-0.074	-0.142	-0.006	0.034	-2.142	0.033
Motion	-0.13	-0.224	-0.036	0.048	-2.719	0.007

Figure 5.1: LIWC attributes to predict the trait neuroticism.



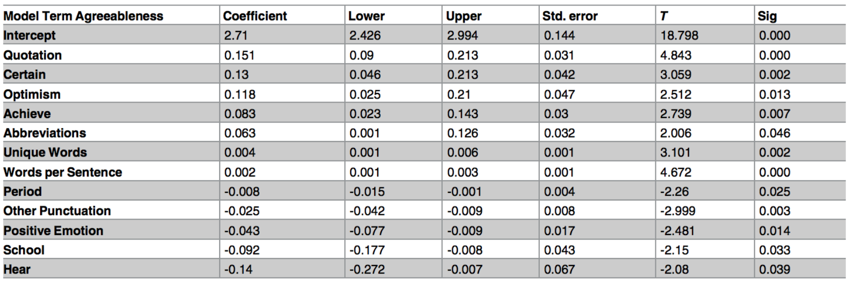
Model Term Conscientiousness	Coefficient	Lower	Upper	Std. error	*T*	Sig
Intercept	3.232	3.002	3.462	0.117	27.717	0.000
Assent	0.449	0.156	0.741	0.149	3.021	0.003
Home	0.258	0.119	0.397	0.071	3.657	0.000
Hear	0.168	0.023	0.313	0.074	2.286	0.023
Occupation	0.068	0.025	0.112	0.022	3.1	0.002
Quotation	0.066	0.003	0.129	0.032	2.052	0.041
Cognitive Mechanisms	0.056	0.021	0.09	0.018	3.16	0.002
Apostrophes	0.044	0.001	0.086	0.022	2.001	0.046
+6 Letters in Words	0.013	0.004	0.021	0.004	2.865	0.005
Recognized by Dictionary	0.01	0.002	0.017	0.004	2.497	0.013
Period	-0.009	-0.016	-0.002	0.003	-2.549	0.011
Self	-0.027	-0.052	-0.003	0.012	-2.22	0.027
Prepositions	-0.028	-0.05	-0.006	0.011	-2.51	0.013
Tentative	-0.076	-0.128	-0.024	0.027	-2.863	0.005
Motion	-0.079	-0.156	-0.003	0.039	-2.043	0.042
Communication	-0.097	-0.163	-0.031	0.033	-2.9	0.004
Optimism	-0.097	-0.18	-0.014	0.042	-2.289	0.023
School	-0.1	-0.195	-0.004	0.048	-2.061	0.04
Family	-0.138	-0.266	-0.011	0.065	-2.135	0.034
Parentheses	-0.139	-0.274	-0.004	0.069	-2.021	0.044
Music	-0.176	-0.349	-0.002	0.088	-1.992	0.047
Inhibition	-0.335	-0.548	-0.123	0.108	-3.108	0.002

Figure 5.2: LIWC attributes to predict the trait conscientiousness.



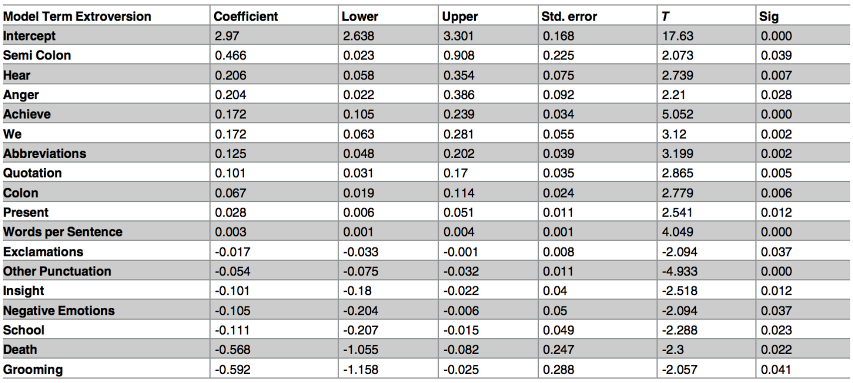
Model Term Openness	Coefficient	Lower	Upper	Std. error	*T*	Sig
Intercept	3.058	2.781	3.334	0.14	21.766	0.000
Anger	0.108	0.006	0.209	0.052	2.086	0.038
Abbreviation	0.077	0.017	0.138	0.031	2.511	0.013
Dashes	0.035	0.005	0.066	0.015	2.316	0.021
Recognized by Dictionary	0.007	0.001	0.013	0.003	2.408	0.017
Words per Sentence	0.001	0	0.003	0.001	2.809	0.005
Inclusion	-0.033	-0.063	-0.002	0.015	-2.107	0.036
Apostrophes	-0.042	-0.083	0	0.021	-1.97	0.05
Discrepancies	-0.078	-0.141	-0.016	0.032	-2.462	0.014
Humans	-0.107	-0.19	-0.024	0.042	-2.524	0.012
Motion	-0.117	-0.191	-0.044	0.037	-3.131	0.002
Semi Colon	-0.458	-0.83	-0.086	0.189	-2.423	0.016

Figure 5.3: LIWC attributes to predict the trait openness.



Model Term Agreeableness	Coefficient	Lower	Upper	Std. error	*T*	Sig
Intercept	2.71	2.426	2.994	0.144	18.798	0.000
Quotation	0.151	0.09	0.213	0.031	4.843	0.000
Certain	0.13	0.046	0.213	0.042	3.059	0.002
Optimism	0.118	0.025	0.21	0.047	2.512	0.013
Achieve	0.083	0.023	0.143	0.03	2.739	0.007
Abbreviations	0.063	0.001	0.126	0.032	2.006	0.046
Unique Words	0.004	0.001	0.006	0.001	3.101	0.002
Words per Sentence	0.002	0.001	0.003	0.001	4.672	0.000
Period	-0.008	-0.015	-0.001	0.004	-2.26	0.025
Other Punctuation	-0.025	-0.042	-0.009	0.008	-2.999	0.003
Positive Emotion	-0.043	-0.077	-0.009	0.017	-2.481	0.014
School	-0.092	-0.177	-0.008	0.043	-2.15	0.033
Hear	-0.14	-0.272	-0.007	0.067	-2.08	0.039

Figure 5.4: LIWC attributes to predict the trait agreeableness.



Model Term Extroversion	Coefficient	Lower	Upper	Std. error	*T*	Sig
Intercept	2.97	2.638	3.301	0.168	17.63	0.000
Semi Colon	0.466	0.023	0.908	0.225	2.073	0.039
Hear	0.206	0.058	0.354	0.075	2.739	0.007
Anger	0.204	0.022	0.386	0.092	2.21	0.028
Achieve	0.172	0.105	0.239	0.034	5.052	0.000
We	0.172	0.063	0.281	0.055	3.12	0.002
Abbreviations	0.125	0.048	0.202	0.039	3.199	0.002
Quotation	0.101	0.031	0.17	0.035	2.865	0.005
Colon	0.067	0.019	0.114	0.024	2.779	0.006
Present	0.028	0.006	0.051	0.011	2.541	0.012
Words per Sentence	0.003	0.001	0.004	0.001	4.049	0.000
Exclamations	-0.017	-0.033	-0.001	0.008	-2.094	0.037
Other Punctuation	-0.054	-0.075	-0.032	0.011	-4.933	0.000
Insight	-0.101	-0.18	-0.022	0.04	-2.518	0.012
Negative Emotions	-0.105	-0.204	-0.006	0.05	-2.094	0.037
School	-0.111	-0.207	-0.015	0.049	-2.288	0.023
Death	-0.568	-1.055	-0.082	0.247	-2.3	0.022
Grooming	-0.592	-1.158	-0.025	0.288	-2.057	0.041

Figure 5.5: LIWC attributes to predict the trait extraversion.

By using these regression models, we have calculated each of the big five personality traits scores. Calculated scores and real scores are showed for each user in Table 5.1.

Table 5.1: LIWC Used Classification

Calc. conscientiousness	**Actual conscientiousness**	**Calc. Openness**	**Actual Openness**	**Calc. extraversion**	**Actual extraversion**	**Calc. Agreeableness**	**Actual Agreeableness**	**Calc. neuroticism**	**Actual neuroticism**
46	65	31	60	25	65	19	60	32	42
42	40	32	71	18	35	18	60	32	85
42	29	31	90	24	81	20	56	31	71
42	71	32	79	20	60	18	62	30	48
42	69	32	77	21	35	19	56	32	8
45	42	31	71	23	48	22	54	32	79
42	50	31	75	21	62	19	62	32	35
44	73	32	77	21	62	21	60	31	37
44	75	31	87	26	79	21	65	30	29
45	46	32	85	22	25	19	50	32	52
44	67	33	71	22	71	24	62	33	60
52	67	31	81	27	71	22	73	31	48
40	60	32	81	15	56	18	73	34	35
43	60	31	83	26	40	20	58	32	54
43	81	28	90	30	92	16	65	28	33
46	46	31	83	21	75	19	33	31	44
46	69	31	90	22	69	19	69	30	71
41	62	33	67	23	79	21	62	32	35
43	56	37	85	6	56	25	46	25	71
42	56	33	71	20	48	18	54	32	62
39	44	32	54	23	65	21	67	33	33
47	52	32	73	21	48	20	54	29	37
45	46	30	79	18	48	17	56	28	48
44	58	32	90	14	67	21	71	34	12
46	50	31	77	23	65	20	69	33	46
41	60	30	71	21	54	19	48	33	46
43	44	32	71	16	62	19	46	31	50
42	58	32	69	21	60	20	54	32	44
47	67	32	67	25	52	15	69	34	23
42	73	31	73	21	87	19	73	30	58
41	44	34	81	18	54	18	56	30	35
45	83	31	75	22	67	20	75	31	17
43	65	32	71	15	75	16	50	30	67

We have evaluated our results by using three metrics as it can be seen in Table 5.2. We had the best results in Conscientiousness and Neuroticism dimension. However, even these two results cannot be considered as accurate.

R-squared score is measure which describes the closeness between the data and the regression line. R-squared score is the variance proportion in the variables which are dependent and predictable from the independent variable. Value one means that the two variables are perfectly correlated. In other words, if r-squared score is equal to one then there is no variance at all. Lower the value means lower correlation. This score is expected to be between zero and one. A low value would show a low level of correlation that means that regression model is not valid. As it can be seen in Table 5.2, our r-squared values are all lower than zero and it means that the regression models we are using cannot be considered as valid.

Mean Absolute Error(MAE), is the absolute errors' mean as it can be understood from its name. MAE tells about the expected error from the forecasted value. The scale we are working on is between zero and one hundred. Thus, our models for openness, extraversion and agreeableness are invalid with the MAE values greater than 40.

Mean Squarred Error(MSE) is the average of the squares of the mean of absolute errors. RMSE is the square root of MSE. It expresses average model prediction error in units just like MAE. However, RMSE penalizes huge errors when it is compared to MAE. In our case, the actual values are up to one hundred. In a big scale like that, huge errors can be detected better with RMSE because the errors are squared before they are averaged. Results of all dimensions seem above the acceptable range.

By using the already proven regression models, we have detected that LIWC features don't show acceptable behaviors when English version of LIWC is used with translated Turkish text data. Thus, we didn't consider LIWC features in our dataset in the classification works.

Table 5.2: Prediction Accuracy of LIWC Features

	R2 Score	**MAE**	**RMSE**
Openness	-27.70	44.81	45.66
Conscientiousness	-1.4	16.21	19.77
Extraversion	-7.04	40.06	42.64
Agreeableness	-18.50	40.15	41.27
Neuroticism	-0.71	19.06	23.83

5.1.2 Using Other Features

To choose the right estimator we have followed the instructions which are shown in Figure 5.6.

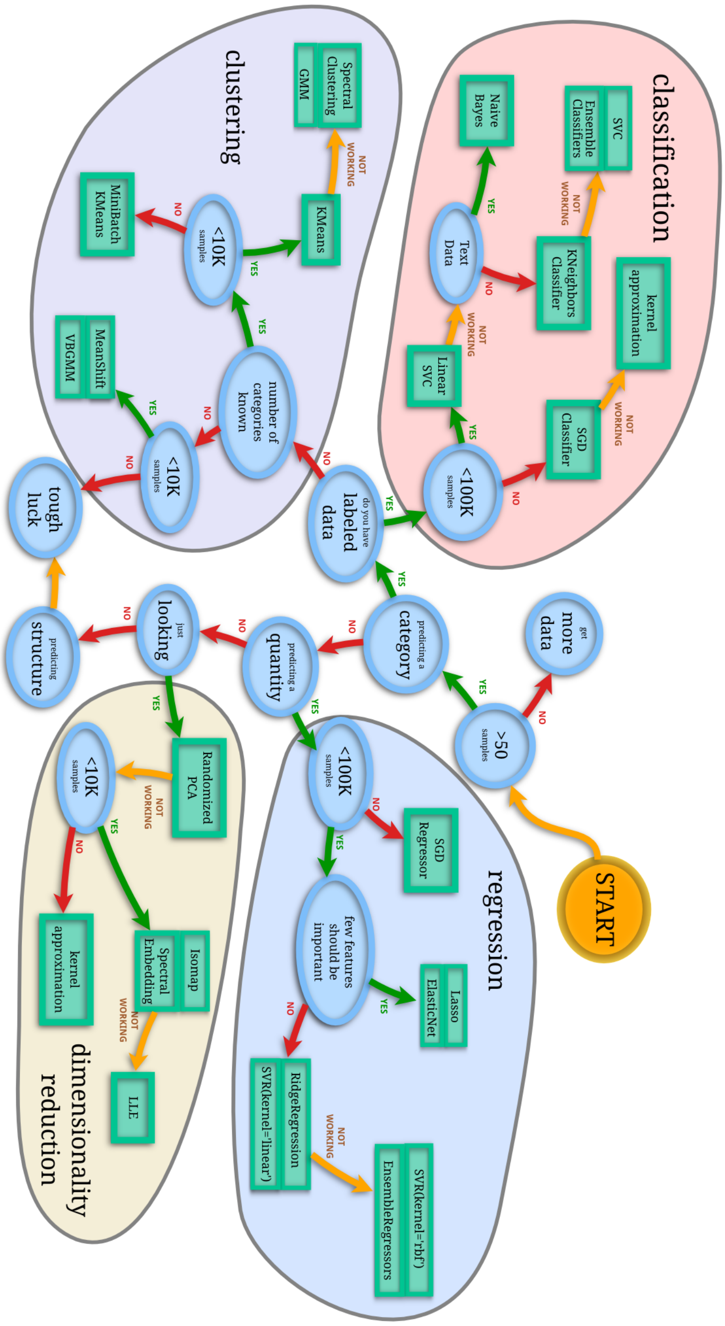


Figure 5.6: Classification Algorithm Selection

Because we had data from 40 volunteers, the possible estimators were SVC, Linear SVC, KNeighbors Classifier, Nearest Centroid Classifier and Ensemble Classifier. As ensemble classifier, we have used Random Forest Classifier.

SVC support vector classification algorithm aims to find a hyperplane in an space to separate the different classes of data points From the many possible hyperplanes SVC finds the optimum hyperplane which has the maximum distance between data points in different classes. By maximizing the margin distance we aim to have reinforcement to classify future data points with more confidence. We have used SVC algorithm of Sci-kit learn with;

- C=1.0,
- kernel='rbf'
- degree=3
- gamma='auto_deprecated'
- coef0=0.0
- shrinking=True
- probability=False
- tol=0.001
- cache_size=200
- max_iter=-

Linear SVC Linear SVC works similar to SVC. Linear SVC has better flexibility in choosing loss functions and penalties. We have used Linear SVC algorithm of Sci-kit learn with;

- penalty='l2'
- loss='squared_hinge'
- dual=True
- tol=0.0001
- C=1.0

- multi_class='ovr'
- fit_intercept=True
- intercept_scaling=1
- class_weight=None
- max_iter=1000

KNeighbors Classifier KNeighbors Classifier classifies an object by a plurality of its neighbors. An object belongs to the class which most common among its K nearest neighbors. For example if K is equal to 1, an object is in the class of that single nearest neighbor. Euclidean distance is most commonly used method to calculate the distance. We have used KNeighbors Classifier algorithm of Sci-kit learn with;

- k=5
- weights='uniform'
- leaf_size=30
- algorithm='auto'
- p=2
- metric='euclidean'

Random Forest Classifier It contains multiple decision trees which works as an ensemble. The one with the highest vote becomes the model's prediction. It averages those decision trees to have better accuracy and to control over fitting. We have used Random Forest Classifier algorithm of Sci-kit learn with;

- min_samples_split=2
- min_samples_leaf=1
- min_weight_fraction_leaf=0.0

- min_impurity_decrease=0.0
- bootstrap=True
- n_estimators='warn'
- criterion='gini'

5.1.3 Evaluation

We have done the evaluation by two methods. One of them is the default accuracy score calculation method of Sci-kit Learn. The other is a generic metric called K-Fold Cross Validation.

Accuracy Score Method Of Sci-kit Learn This method is created to be used in multi label classification. It computes subset accuracy. In other words, the set of labels predicted for a sample must exactly match the corresponding ground truth labels. We have used this method of Sci-kit learn with;

- normalize=True,
- ample_weight=None

K-Fold Cross Validation To evaluate the classification model used, we generally divide the data into two namely training and test sets. However, different divisions cause different results and this causes a problem to evaluate the model. K-Fold Cross Validation solves this problem by dividing the data into folds. This metric ensures that each fold is used as both testing and training set. The number K here defines the number of sections that the dataset will be divided.

Our evaluation scores are shown in Table 5.3, Table 5.4, Table 5.5, Table 5.6, Table 5.7.

Table 5.3: Prediction Accuracy for Openness

	Accuracy Score	**3-Fold Cross Validation**
Random Forest Classifier	1.0	0.54 (+/- 0.18)
SVC	0.66	0.67 (+/- 0.05)
LinearSVC	0.66	0.47 (+/- 0.05)
KNeighbors Classifier	0.4	0.44 (+/- 0.11)

Table 5.4: Prediction Accuracy for Conscientiousness

	Accuracy Score	**3-Fold Cross Validation**
Random Forest Classifier	0.96	0.46 (+/- 0.28)
SVC	0.53	0.53 (+/- 0.05)
LinearSVC	0.66	0.5 (+/- 0.14)
KNeighbors Classifier	0.5	0.42 (+/- 0.34)

Table 5.5: Prediction Accuracy for Extraversion

	Accuracy Score	**3-Fold Cross Validation**
Random Forest Classifier	0.96	0.32 (+/- 0.31)
SVC	0.53	0.54 (+/- 0.12)
LinearSVC	0.76	0.62 (+/- 0.39)
KNeighbors Classifier	0.43	0.28 (+/- 0.33)

Table 5.6: Prediction Accuracy for Agreeableness

	Accuracy Score	**3-Fold Cross Validation**
Random Forest Classifier	0.96	0.83 (+/- 0.09)
SVC	0.80	0.8 (+/- 0.07)
LinearSVC	0.90	0.83 (+/- 0.09)
KNeighbors Classifier	0.80	0.77 (+/- 0.09)

Table 5.7: Prediction Accuracy for Neuroticism

	Accuracy Score	**3-Fold Cross Validation**
Random Forest Classifier	0.96	0.57 (+/- 0.05)
SVC	0.63	0.63 (+/- 0.05)
LinearSVC	0.63	0.57 (+/- 0.24)
KNeighbors Classifier	0.3	0.17 (+/- 0.1)

5.1.4 Classification With English Text

To make sure that TF-IDF weighting and word2vec embedding creates features which are valid for personality detection from social media content, we have used an already created English dataset [1]. This dataset has been used in multiple researches and accurate results have been procured by using it.

This dataset has Facebook status updates of 251 users. Also, it contains big five personality scores of the users. In this dataset, each post of the user is given in a different line. Thus, firstly we needed to gether textual data of each user. Then, we have used this dataset to test the features which are found by word2vec embedding. The phases followed are the same as we have used for Turkish textual data.

[1] https://github.com/Myoungs/myPersonality-dataset/blob/master/mypersonality.csv

Normalization and Stemming Firstly, we have cleaned the data from double spaces and punctuation marks to ensure a textual data ready for stemming. Then, by using Porter Stemmer we have replaced the words by their lemmas.

TF-IDF Weighting By using Scikit-learn's TF-IDF counter, we have found the mostly used words for each user. In order to do this;

- We have used a stop word dictionary to not consider them while calculating TF-IDF scores.
- Scikit-learn's algorithm is arranged to eliminate the words whose DF scores are more than 0.8. By doing it, we have eliminated users who doesn't have enough posts.
- We have found not only the single words but also 2-grams and 3-grams.

After sorting the words for each user according to their TF-IDF scores, we have calculated the top used 20 words for each user.

Word2Vec Embedding To create feature vector for each user by using top used 20 words we have used a ready to use word embedding which has been created from wikipedia[2]. It represents each word by 300 features. We have summed the vectors of each word and ended up with 300 features for each user.

Classification For the classification we have used 4 different algorithms which are explained in section Using Other Features.

Evaluation To evaluate the results we have found by classification, we have used 3-Fold cross validation and accuracy score as it is described in section Evaluation. The comparison of accuracies of Turkish and English word2vec features are shown in Table 5.8, Table 5.9, Table 5.10, Table 5.11, Table 5.12.

[2] https://www.kaggle.com/yesbutwhatdoesitmean/wikinews300d1mvec

Table 5.8: Prediction Accuracy for Openness by Word2Vec Features

	Accuracy Score (English)	**Accuracy Score (Turkish)**	**3-Fold Cross Validation (English)**	**3-Fold Cross Validation (Turkish)**
Random Forest Classifier	0.98	0.93	0.43 (+/- 0.09)	0.56 (+/- 0.24)
SVC	0.49	0.67	0.5 (+/- 0.02)	0.67 (+/- 0.05)
LinearSVC	1.0	0.66	0.42 (+/- 0.07)	0.67 (+/- 0.05)
KNeighbors Classifier	0.50	0.59	0.36 (+/- 0.19)	0.46 (+/- 0.28)

Table 5.9: Prediction Accuracy for Conscientiousness by Word2Vec Features

	Accuracy Score (English)	**Accuracy Score (Turkish)**	**3-Fold Cross Validation (English)**	**3-Fold Cross Validation (Turkish)**
Random Forest Classifier	0.97	0.93	0.47 (+/- 0.17)	0.52 (+/- 0.29)
SVC	0.61	0.63	0.44 (+/- 0.07)	0.63 (+/- 0.08)
LinearSVC	1.0	0.63	0.42 (+/- 0.09)	0.63 (+/- 0.08)
KNeighbors Classifier	0.57	0.46	0.37 (+/- 0.09)	0.44 (+/- 0.17)

Table 5.10: Prediction Accuracy for Extraversion by Word2Vec Features

	Accuracy Score (English)	**Accuracy Score (Turkish)**	**3-Fold Cross Validation (English)**	**3-Fold Cross Validation (Turkish)**
Random Forest Classifier	0.99	0.96	0.47 (+/- 0.05)	0.53 (+/- 0.3)
SVC	0.66	0.53	0.51 (+/- 0.1)	0.53 (+/- 0.05)
LinearSVC	1.0	0.53	0.5 (+/- 0.09)	0.53 (+/- 0.05)
KNeighbors Classifier	0.66	0.43	0.44 (+/- 0.13	0.46 (+/- 0.18)

Table 5.11: Prediction Accuracy for Agreeableness by Word2Vec Features

	Accuracy Score (English)	**Accuracy Score (Turkish)**	**3-Fold Cross Validation (English)**	**3-Fold Cross Validation (Turkish)**
Random Forest Classifier	0.99	0.96	0.46 (+/- 0.07)	0.5 (+/- 0.43)
SVC	0.55	0.8	0.53 (+/- 0.05)	0.8 (+/- 0.12)
LinearSVC	1.0	0.8	0.48 (+/- 0.04)	0.8 (+/- 0.12)
KNeighbors Classifier	0.5	0.63	0.37 (+/- 0.08)	0.57 (+/- 0.26)

Table 5.12: Prediction Accuracy for Neuroticism by Word2Vec Features

	Accuracy Score (English)	**Accuracy Score (Turkish)**	**3-Fold Cross Validation (English)**	**3-Fold Cross Validation (Turkish)**
Random Forest Classifier	0.99	0.90	0.69 (+/- 0.07)	0.42 (+/- 0.24)
SVC	0.74	0.53	0.75 (+/- 0.02)	0.54 (+/- 0.12)
LinearSVC	1.0	0.56	0.62 (+/- 0.11)	0.54 (+/- 0.12)
KNeighbors Classifier	0.64	0.36	0.48 (+/- 0.2)	0.35 (+/- 0.24)

Results From the evaluation scores we have deduced that;

- Bigger sample size gives smaller standard deviation. This showed us our framework would give lower std values if we have trained it with bigger sample size.
- The results we have found by using Turkish social media content gives similar accuracy results.
- SVC gives better results compared to the other classification algorithms. The same implication is valid for the classification we have done by Turkish content.

CHAPTER 6

CONCLUSIONS

In this thesis, we have presented a methodology to predict personality traits by using Twitter Data. We have designed our framework to work in Turkish tweets. However, the whole process can be done in any language with different tools designed for that language.

After collecting the data, firstly we have done preprocessing on it. In this phase, we have eliminated some of the tweets. Then cleaned the tweets from Twitter specific usages and the noise which are coming from informal usage of Twitter. By using the cleaned data, we have selected our features. Some of the features were coming from a tool designed for English texts called LIWC. To utilize this tool, we needed to translate the text data into English. Then by using NLP techniques, we have determined linguistic features including the emoticon selection. We have also used tweet related features such as timestamp. After determining these features, we have eliminated useless ones and normalized the rest of it. Finally, this eliminated feature set has been concatenated with the word vector which we have adjusted by Word2Vec based word embedding. Because LIWC is not designed for Turkish, we have checked if it works with Turkish text by using a already proven regression models. We have seen that translating the tweets causes LIWC to work wrong. Because of this, we haven't used the features coming from LIWC. The classification has been done by the rest of the features with different algorithms. Because we had a relatively small number of data as ground truth, we couldn't prove the validity of our framework and proving the validity of our framework with a bigger ground truth set remains as an open issue.

REFERENCES

[1] P. Rozin, “Social psychology and science: Some lessons from solomon asch,” *Personality and Social Psychology Review - PERS SOC PSYCHOL REV*, vol. 5, pp. 2–14, 02 2001.

[2] J. A Bargh and K. Y A McKenna, “The internet and social life,” *Annual review of psychology*, vol. 55, pp. 573–90, 02 2004.

[3] D. Suhartono, V. Ong, A. D. S. Rahmanto, N. given name Williem, A. E. Nugroho, E. Andangsari, and M. Suprayogi, “Personality prediction based on twitter information in bahasa indonesia,” pp. 367–372, 09 2017.

[4] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, “Social media mobile internet use among teens and young adults,” *Pew Internet and American Life Project*, 01 2010.

[5] J. Raacke and J. Bonds-Raacke, “Myspace and facebook: Applying the uses and gratifications theory to exploring friend-networking sites,” *Cyberpsychology behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, vol. 11, pp. 169–74, 05 2008.

[6] J. Ann Golbeck, “Computing and applying trust in web-based social networks,” 01 2005.

[7] J. Baddeley and J. Singer, “A loss in the family: Silence, memory, and narrative identity after bereavement,” *Memory (Hove, England)*, vol. 18, pp. 198–207, 09 2009.

[8] L. Fast and D. Funder, “Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior,” *Journal of personality and social psychology*, vol. 94, pp. 334–46, 03 2008.

[9] J. Hirsh and J. Peterson, “Extraversion, neuroticism, and the prisoner’s

dilemma," *Personality and Individual Differences*, vol. 46, pp. 254–256, 12 2009.

[10] J. Pennebaker and L. King, "Linguistic styles: Language use as an individual difference," *Journal of personality and social psychology*, vol. 77, pp. 1296–312, 01 2000.

[11] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, pp. 547–77, 02 2003.

[12] M. Mehl, S. Gosling, and J. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," *Journal of personality and social psychology*, vol. 90, pp. 862–77, 06 2006.

[13] N. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students' use of online social network sites," *J. Computer-Mediated Communication*, vol. 12, pp. 1143–1168, 07 2007.

[14] K. Y. A. McKenna, A. S. Green, and M. Gleason, "Relationship formation on the internet: What's the big attraction?," *Journal of Social Issues - J SOC ISSUES*, vol. 58, pp. 9–31, 01 2002.

[15] A. Manago, M. B. Graham, P. Greenfield, and G. Salimkhan, "Self-presentation and gender on myspace," *Journal of Applied Developmental Psychology*, vol. 29, pp. 446–458, 09 2008.

[16] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, 03 2013.

[17] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.

[18] J. Zywica and J. Danowski, "The faces of facebookers: Investigating social enhancement and social compensation hypotheses; predicting facebook™ and offline popularity from sociability and self-esteem, and mapping the meanings of

popularity with semantic networks," *Journal of Computer-Mediated Communication*, vol. 14, pp. 1 – 34, 11 2008.

[19] P. Rosen and D. Kluemper, "The impact of the big five personality traits on the acceptance of social networking website," *Peter A. Rosen*, vol. 2, 01 2008.

[20] K. Wilson, S. Fornasier, and K. M White, "Psychological predictors of young adults' use of social networking sites," *Cyberpsychology, behavior and social networking*, vol. 13, pp. 173–7, 04 2010.

[21] T. Ryan and S. Xenos, "Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage," *Computers in Human Behavior*, vol. 27, pp. 1658–1664, 09 2011.

[22] T. Correa, A. Willard Hinsley, and H. Gil de Zúñiga, "Who interacts on the web?: The intersection of users' personality and social media use," *Computers in Human Behavior*, vol. 26, pp. 247–253, 03 2010.

[23] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with facebook use," *Computers in Human Behavior*, vol. 25, pp. 578–586, Mar. 2009.

[24] Y. Amichai-Hamburger, "Internet and personality," *Computers in Human Behavior*, vol. 18, pp. 1–10, 01 2002.

[25] S. Gosling, A. A Augustine, S. Vazire, N. Holtzman, and S. Gaddis, "Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information," *Cyberpsychology, behavior and social networking*, vol. 14, pp. 483–8, 09 2011.

[26] K. Moore and J. Mcelroy, "The influence of personality on facebook usage, wall postings, and regret," *Computers in Human Behavior*, vol. 28, pp. 267–274, 01 2012.

[27] R. Guadagno, B. Okdie, and C. A. Eno, "Who blogs? personality predictors of blogging," *Computers in Human Behavior*, pp. 1993–2004, 05 2013.

[28] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and patterns of facebook usage," *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci'12*, 06 2012.

[29] J. Schrammel, C. Köffel, and M. Tscheligi, "Personality traits, usage patterns and information disclosure in online communities," in *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, BCS-HCI '09, (Swinton, UK, UK), pp. 169–174, British Computer Society, 2009.

[30] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Computers in Human Behavior*, vol. 26, pp. 1289–1295, 11 2010.

[31] A. S. Acar and M. Polonsky, "Online social networks and insights into marketing communications," *Journal of Internet Commerce*, vol. 6, no. 4, pp. 55–72, 2007.

[32] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," pp. 180–185, 10 2011.

[33] J. Asendorpf and S. Wilpers, "Personality effects on social relationships," *Journal of Personality and Social Psychology*, vol. 74, pp. 1531–1544, 06 1998.

[34] M. Zalk, W. Burk, S. Branje, J. Denissen, M. Aken, and W. Meeus, "Emerging late adolescent friendship networks and big five personality traits: A social network approach," *Journal of personality*, vol. 78, pp. 509–38, 04 2010.

[35] F. Celli, "Mining user personality in twitter," 05 2019.

[36] C. Barrett, *Pew Internet and American Life Project*, pp. 1464–1465. 01 2013.

[37] G. Park, H. Schwartz, J. Eichstaedt, M. Kern, M. Kosinski, D. Stillwell, L. Ungar, and M. E P Seligman, "Automatic personality assessment through social media language," *Journal of personality and social psychology*, vol. 108, 11 2014.

[38] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text.," *J. Artif. Intell. Res. (JAIR)*, vol. 30, pp. 457–500, 09 2007.

[39] J. Golbeck, C. Robles, and K. Turner, “Predicting personality with social media,” pp. 253–262, 01 2011.

[40] S. Counts and K. Brooke Stecher, “Self-presentation of personality during online profile creation.,” 01 2009.

[41] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu, “Improving user profile with personality traits predicted from social media content,” pp. 355–358, 10 2013.

[42] D. Wan, C. Zhang, M. Wu, and Z. An, “Personality prediction based on all characters of user social media information,” vol. 489, pp. 220–230, 11 2014.

[43] K.-H. Peng, L.-H. Liou, C.-S. Chang, and D.-S. Lee, “Predicting personality traits of chinese users based on facebook wall posts,” pp. 9–14, 10 2015.

[44] M. Arroju, A. Hassan, and G. Farnadi, “Age, gender and personality recognition using tweets in a multilingual setting,” in *CLEF 2015 working notes*, 2015.

[45] B. Yudha Pratama and R. Sarno, “Personality classification based on twitter text using naive bayes, knn and svm,” pp. 170–174, 11 2015.

[46] M. Tsytsarau and T. Palpanas, “Mining subjective data on the web,” *Data Mining and Knowledge Discovery*, vol. 24, pp. 478–514, 05 2011.

[47] A. C. Lima and L. De Castro, “Multi-label semi-supervised classification applied to personality prediction in tweets,” 09 2013.

[48] A. Byers, R. Boochever, C. Sumner, A. Byers, R. Boochever, and G. J Park, “Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets,” *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, vol. 2, pp. 383–396, 12 2012.

[49] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. M Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E P Seligman, and L. Ungar, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PloS one*, vol. 8, p. e73791, 09 2013.

[50] Y. Liu, J. Wang, and Y. Jiang, “Pt-lda: A latent variable model to predict personality traits of social network users,” *Neurocomputing*, vol. 210, 06 2016.

[51] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, “Recognising personality traits using facebook status updates,” *AAAI Workshop - Technical Report*, pp. 14–18, 01 2013.

[52] P. Bibby, “Dispositional factors in the use of social networking sites: Findings and implications for social computing research,” vol. 5075, pp. 392–400, 06 2008.

[53] O. P. John, J. Cheek, and E. C. Klohnen, “On the nature of self-monitoring: Construct explication with q-sort ratings,” *Journal of personality and social psychology*, vol. 71, pp. 763–76, 11 1996.

[54] P. Trapnell and J. Campbell, “Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection,” *Journal of Personality and Social Psychology*, vol. 76, pp. 284–304, 02 1999.

[55] A. L. Forest and J. V. Wood, “When social networking is not working: Individuals with low self-esteem recognize but do not reap the benefits of self-disclosure on facebook,” *Psychological Science*, vol. 23, no. 3, pp. 295–302, 2012.

[56] M. Hall and S. Caton, “Am i who i say i am? unobtrusive self-representation and personality recognition on facebook,” *PLOS ONE*, vol. 12, p. e0184417, 09 2017.

[57] J. Oberlander and S. Nowson, “Whose thumb is it anyway?: Classifying author personality from weblog text,” in *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL ’06, (Stroudsburg, PA, USA), pp. 627–634, Association for Computational Linguistics, 2006.

[58] T. Yarkoni, “Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers,” *Journal of research in personality*, vol. 44, pp. 363–373, 06 2010.

[59] F. Iacobelli, A. Gill, S. Nowson, and J. Oberlander, “Large scale personality classification of bloggers,” in *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, pp. 568–577, Springer-Verlag GmbH, 2011.

[60] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock, and S. Davalos, “A multivariate regression approach to personality impression recognition of vloggers,” *WCPR 2014 - Proceedings of the 2014 Workshop on Computational Personality Recognition, Workshop of MM 2014*, pp. 1–6, 11 2014.

[61] J. Stoughton, L. Foster Thompson, and A. Meade, “Big five personality traits reflected in job applicants’ social media postings,” *Cyberpsychology, behavior and social networking*, 06 2013.

[62] D. Hughes, M. Rowe, M. Batey, and A. Lee, “A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage,” *Computers in Human Behavior*, vol. 28, pp. 561–569, 03 2012.

[63] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, “Computational personality recognition in social media,” *User Modeling and User-Adapted Interaction*, vol. 26, 02 2016.

[64] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, “Fusing social media cues: Personality prediction from twitter and instagram,” pp. 107–108, 04 2016.

[65] A. Langmeyer, C. Tarnai, and A. Guglhör-Rudan, “What is your music preference telling us about your personality?,” 07 2008.

[66] P. Rentfrow and S. Gosling, “The do re mi’s of everyday life: The structure and personality correlates of music preferences,” *Journal of personality and social psychology*, vol. 84, pp. 1236–56, 07 2003.

[67] G. Odekerken, K. De Wulf, and P. Schumacher, “Strengthening outcomes of retailer–consumer relationships: The dual impact of relationship marketing tactics and consumer personality,” *Journal of Business Research*, pp. 177–190, 02 2003.

[68] R. Boele, *The Big Five Personality Factors: The psycholexical approach to personality.* 01 2000.

[69] L. R. Goldberg, “An alternative "description of personality": The big-five factor structure.,” *Journal of Personality and Social Psychology*, vol. 59, no. 6, pp. 1216–1229, 1990.

[70] L.-F. Zhang, “Thinking styles and the big five personality traits,” *Educational Psychology - EDUC PSYCHOL-UK*, vol. 22, pp. 17–31, 01 2002.

[71] M. Ashton, K. Lee, and S. V Paunonen, “What is the central feature of extraversion? social attention versus reward sensitivity,” *Journal of personality and social psychology*, vol. 83, pp. 245–52, 08 2002.

[72] P. Costa and T. A. Widiger, “Personality disorders and the five - factor model of personality,” 01 2012.

[73] W. Graziano and N. Eisenberg, *Agreeableness*, pp. 795–824. 12 1997.

[74] J. A. LEPINE, J. Colquitt, and A. Erez, “Adaptability to changing task contexts: Effects of general cognitive ability conscientiousness, and openness to experience,” *Personnel Psychology*, vol. 53, pp. 563 – 593, 12 2006.

[75] R. R. McCrae and P. Costa, “Personality trait structure as a human universal,” *The American psychologist*, vol. 52, pp. 509–16, 06 1997.

[76] D. G. Winter, O. P. John, A. J. Stewart, E. C. Klohnen, and L. Duncan, “Traits and motives: Toward an integration of two traditions in personality research,” *Psychological review*, vol. 105, pp. 230–50, 05 1998.

[77] D. A. Cobb-Clark and S. Schurer, “The stability of big-five personality traits,” *Economics Letters*, vol. 115, 09 2011.

[78] D. Schmitt, J. Allik, R. R. McCrae, V. Benet, L. Alcalay, L. Ault, I. Austers, K. Bennett, G. Bianchi, F. Boholst, M. Ann Borg Cunen, J. Braeckman, E. G. Brainerd Jr, L. G. Caral, G. Caron, M. Martina Casullo, M. Cunningham, I. Daibo, C. de backer, and A. Zupanèiè, “The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations,” *Journal of Cross-Cultural Psychology*, vol. 38, pp. 173–212, 01 2007.

[79] J. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count (liwc): Liwc2001,” vol. 71, 01 2001.

[80] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words:

Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[81] C.-L. Huang, C. Chung, N. Hui, Y.-C. Lin, Y.-T. Seih, B. Lam, and J. Pennebaker, "Development of the chinese linguistic inquiry and word count dictionary," *Chinese J Psychol*, vol. 54, pp. 185–201, 01 2012.

[82] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text mining: Techniques, applications and issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, 11 2016.

[83] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, KDD'95, pp. 112–117, AAAI Press, 1995.

[84] S. Salloum, M. Al-Emran, A. Abdel Monem, and K. Shaalan, "A survey of text mining in social media: Facebook and twitter perspectives," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, pp. 127–133, 01 2017.

[85] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37–54, 03 1996.

[86] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, (Stroudsburg, PA, USA), pp. 3–10, Association for Computational Linguistics, 1999.

[87] A. Oulasvirta, E. Lehtonen, E. Kurvinen, and M. Raento, "Making the ordinary visible in microblogs," *Personal and Ubiquitous Computing*, vol. 14, pp. 237–249, 04 2010.

[88] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," *of the 9th WebKDD and 1st SNA*, vol. 43, pp. 56–65, 01 2007.

[89] P. McFedries, "Technically speaking: All a-twitter," *Spectrum, IEEE*, vol. 44, pp. 84 – 84, 11 2007.

[90] A. Nadkarni and S. Hofmann, “Why do people use facebook?,” *Personality and individual differences*, vol. 52, pp. 243–249, 02 2012.

[91] S. Zhao, S. Grasmuck, and J. Martin, “Identity construction on facebook: Digital empowerment in anchored relationships,” *Computers in Human Behavior*, vol. 24, pp. 1816–1836, 09 2008.

[92] S. Utz, M. Tanis, and I. Vermeulen, “It is all about being popular: The effects of need for popularity on social network site use,” *Cyberpsychology, Behavior, and Social Networking*, vol. 15, pp. 37–42, 01 2012.

[93] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling, “Facebook profiles reflect actual personality, not self-idealization,” *Psychological science*, vol. 21, pp. 372–4, 03 2010.

[94] R. Baumeister and M. Leary, “The need to belong: Desire for interpersonal attachments as a fundamental human motivation,” *Psychological bulletin*, vol. 117, pp. 497–529, 06 1995.

[95] K. Sheldon, N. Abad, and C. Hinsch, “A two-process view of facebook use and relatedness need-satisfaction: Disconnection drives use, and connection rewards it,” *Journal of personality and social psychology*, vol. 100, pp. 766–75, 04 2011.

[96] A. Chan and A. A. Freitas, “A new ant colony algorithm for multi-label classification with applications in bioinformatics,” in *2006 Genetic and Evolutionary Computation Conference* (M. Keijzer, ed.), vol. 1, (New York, New York (USA)), pp. 27–34, ACM Press, July 2006.

[97] B. Anderson, P. Fagan, T. Woodnutt, and T. Chamorro-Premuzic, “Facebook psychology: Popular questions answered by research,” *Psychology of Popular Media Culture*, vol. 1, pp. 23–37, 01 2012.

[98] P. Costa and R. R. McCrae, “The five-factor model, five-factor theory, and interpersonal psychology,” *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*, pp. 91–104, 03 2012.