MEF UNIVERSITY

UNDERSTANDING EMOTION FLUCTUATIONS USING SOCIAL MEDIA

Capstone Project

Serkan Ceran

İSTANBUL, 2017

MEF UNIVERSITY

UNDERSTANDING EMOTIONAL FLUCTUATIONS USING SOCIAL MEDIA

Capstone Project

Serkan Ceran

Advisor: Asst. Prof. Ezgi Akpınar

İSTANBUL, 2017

MEF UNIVERSITY

Name of the project: Understanding Emotional Fluctuations Using Social Media
Name/Last Name of the Student: Serkan Ceran
Date of Thesis Defense: 13/10/2017

I hereby state that the graduation project prepared by Serkan Ceran has been completed under my supervision. I accept this work as a "Graduation Project".

13/10/2017
Asst. Prof. Ezgi Akpınar

I hereby state that I have examined this graduation project by Serkan Ceran which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

13/10/2017

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of __________ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Ezgi Akpınar	………………………..
2. Özgür Özlük	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Serkan Ceran	11/09/2017	

EXECUTIVE SUMMARY

UNDERSTANDING EMOTIONAL FLUCTUATIONS USING SOCIAL MEDIA

Serkan Ceran

Advisor: Asst. Prof. Ezgi Akpınar

OCTOBER, 2017, 27 pages

During the last years, the importance of social media is increasing in an amazing way. In this paper, we looked at one such popular microblog platform called Twitter and build models for classifying “tweets” into some specific emotion. We used Turkey’s twitter data in order to explore the change in emotions over time using sentiment analysis. Using LIWC dictionary database, we conducted an emotion analysis of approximately 2.2 million tweets.

We tracked how emotions evolve over time based on the prominent events in and or related to Turkey. Our results showed that there is a significant relationship between emotions and prominent events. We also analyzed the correlation between these emotions and the dollar exchange and made a predictive modeling experiment.

Key Words: Sentiment Analysis, Text Mining, Turkish Text, Emotion Analysis, Prediction

ÖZET

UNDERSTANDING EMOTIONAL FLUSTUATIONS USING SOCIAL MEDIA

Serkan Ceran

Tez Danışmanı: Asst. Prof. Ezgi Akpınar

EKİM, 2017, 27 sayfa

Son yıllarda sosyal medyanın önemi inanılmaz bir şekilde artmaktadır. Bu çalışmada Twitter gibi popüler bir mikroblog platformuna baktık ve "tweet'leri" belirli bazı duygularla eşleştirmeye yönelik modeller oluşturduk. Duyguların zaman içerisinde değişimlerini incelemek için Türkiye'de ki twitter verilerini kullandık. Duygu analizini gerçekleştirmek için LIWC sözlük veritabanı kullandık. Yaklaşık olarak 2.2 milyon tweet datasını analiz ederek duygu analizini gerçekleştirdik.

Duyguları, Türkiye'de ve / veya Türkiye ile ilgili önemli olaylara dayanarak nasıl bir tepkinin geliştiğini izledik. Bulgularımız, duygularla bazı belirgin olaylar arasında anlamlı bir ilişki olduğunu gösterdi. Ayrıca bu duygular ile dolar değişimi arasındaki korelasyonu da analiz ederek tahminsel bir modelleme denemesi gerçekleştirdik.

Anahtar Kelimeler: Düşünce Çözümleme, Metin Madenciliği, Türkçe Metin, Duygu Analizi, Tahminleme

TABLE OF CONTENTS

.

1. INTRODUCTION

In recent times, the popularity of social media has grown exponentially and is increasingly being used as a channel for mass communication, such that the brands consider it as a medium of promotion and people largely use it for content sharing. With the increase in the number of users online, the data generated has increased many folds, bringing in the huge scope for gaining insights into the untapped gold mine, the social media data.

Social media is a way of communication using online tools such as Twitter, Facebook, LinkedIn, and so on. Andreas Kaplan and Michael Haenlein define social media as follows:

"A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content"

Social media spans lots of Internet-based platforms that facilitate human emotions such as:

•Networking, for example, Facebook, LinkedIn, and so on

•Micro blogging, for example, Twitter, Tumblr, and so on

•Photo sharing, for example, Instagram, Flickr, and so on

•Video sharing, for example, YouTube, Vimeo, and so on

•Stack exchanging, for example, Stack Overflow, Github, and so on

•Instant messaging, for example, Whatsapp, Hike, and so on

The traditional media such as radio, newspaper, or television, facilitates one-way communication with a limited scope of reach and usability. Though the audience can interact (two-way communication) with these channels, particularly radio, the quality and frequency of such communications are very limited. On the other hand, Internet-based social media offers multi-way communication with features such as immediacy and permanence. It is important to understand all the aspects of social media today because real customers are using it.

In this paper, we look at one such popular microblog platform called Twitter and build models for classifying “tweets” into some specific emotion. Twitter is one of the well-known social media websites, founded in March 21, 2006. It allows users to follow other user’s accounts. Users are able to share what is called tweets up to 140 characters. Nowadays, due to the huge amount of information embedded inside in tweets, different analysis in various areas and sectors have been accomplished and the results are used by different companies or businesses.

In this research, we used Turkey’s twitter data in order to explore change in emotions over time using sentiment analysis. The data is provided from the company named “Crimson Hexagon”. Crimson Hexagon is a social media analytics company, which is an official Twitter partner, that provides raw data based on a random sampling procedure.

1.1. What is Sentiment Analysis?

Sentiment analysis is the process of identifying and categorizing linguistic content in order to understand common patterns such as emotions, attitudes. The term "sentiment" is related to feelings, attitudes, emotions and opinions which are not facts but subjective impressions.

The way people use words in their daily lives can provide rich information about their beliefs, fears, thinking patterns, social relationships, and personalities. From the time of Freud’s writings about slips of the tongue to the early days of computer-based text analysis, researchers began gathering increasingly compelling evidence that the words we use have tremendous psychological value [1-2].

Emotion classification of texts is addressing the problem of defining and categorizing the emotions that can be expressed through language. The specific emotions that be explored in this study is based on LIWC dictionary (Pennebaker, Chung, Ireland, Gonzales and Booth, 2007), which is validated across several languages and types of text.

1.2. Theoretical Background

In very simple terms, conducting sentiment analysis to reveal the emotions in a text is a classification problem. In order to classify text and extract meaningful emotions, there has been variety of different methods have been used such as machine learning (Munmun De Choudhury et al[21].) and word frequency count methods (Mohammed ALSADI et al.[20]).

Sentiment analysis has been used in diverse areas of research including language change over time, mental health, movie sales, politics or even reporting natural disasters. Sentiment analysis might provide description of a macro behavior. For instance, Cavazos-Rehg et al. [7] examined depression-related content in Twitter to glean insight into social networking about mental health.

Sentiment analysis might be used to predict future behavior in various domains.Nguyen et al. [5] have used sentiment analysis to predict the perception of future tweets based on the change previous tweets' perception. Hashtags have been used to capture fine emotion categories from tweets (Mohammad and Kiritchenko, 6). Asur and Huberman [8] used Twitter to forecast box-office revenues of movies. They showed that a simple model built from the rate at which tweets are created about particular topics could outperform market-based predictors. There is also prior work on analyzing correlation between web buzz and stock market. Antweiler and Frank (2004) determine correlation between activity in Internet message boards and stock volatility and trading volume. Other researchers employed blog posts to predict stock market behavior. Gilbert and Karahalios (2010) used over 20 million posts from the LiveJournal website to create an index of the US national mood, which they call the Anxiety Index. They found that when this index rose sharply, the S&P 500 ended the day marginally lower than is expected [13].

Zhang et al. [14] analyzed the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). They looked at whether Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time.

Sentiment analysis has been increasingly used in understanding macro behaviors such as politics. For instance, Tumasjan et al. [9] analyzed Twitter messages mentioning parties and politicians prior to the German federal election 2009 and found that the mere number of tweets reflects voter preferences and comes close to traditional election polls. Finally, sentiment analysis can even predict important events such as natural disasters. Shendge et al. [10] worked on real-time Tweet analysis for event detection and reporting system for Earthquake. While sentiment analysis has been conducted in several languages, use of Turkish texts have been limited. Turkish is a language a Turkic language with about 70 million speakers mainly in Turkey and Cyprus. Although, there are some other countries where the language is spoken, the majority of the Turkish speakers live in Turkey. This make us easier to investigate the emotions that we associate with significant events in Turkey. For instance, image we study emotions using English, the content might have been written due to several events that occur in several regions. Thus, the usage of Turkish written datasets allows us to cleanly investigate emotions that have been written based on specific events that occurred in Turkey. Şimşek and Özdemir[11], used Turkish tweet dataset to determine emotional words. An analysis is carried out to see if there is a relation between Turkish tweets and the Turkish stock market index. To the best of our knowledge, this is the first study performed on Turkish tweets and stock market index. Further, Boynukalın [12] explored Emotion Analysis of Turkish texts. This is the first study that attempts to make a classification of emotions using Turkish texts.

2. EMPIRICAL DATA

In this study, we explored the usage of emotions in language based on the actual events (e.g., terrorist attacks, political changes, prominent events). We have used Twitter data and followed a word count frequency method using the well-established Linguistic Inquiry and Word Count 2007 (LIWC) database in Turkish. Properties of datasets and the methods used to construct them are explained in this section.

2.1. Twitter Data

We have collected tweets written in Turkish for a 242-day period between 16.11.2016 and 15.07.2017. Our data set includes approximately 2.2 million tweets. The daily data is provided from the company named "Crimson Hexagon". There is around 7M Turkish tweets each day. We randomly sampled about 0.12% of the daily tweets in Turkey, which is around 9000 tweets per day.

2.2. Linguistic Inquiry and Word Count 2007 (LIWC) Database

In order to provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples, Pennebaker et al. developed a text analysis application called Linguistic Inquiry and Word Count, or LIWC [15].

The LIWC dictionary contains word-to-category mappings for around 85 categories of words, including both common content words (e.g., words about family, emotions, biological processes) and function words (e.g., pronouns, conjunctions, articles, etc.). For example, the "cognitive processes" category contains words like "think", "understand", and "analyze", and the "articles" category contains the words "a", "an", and "the". Similar techniques have been used to translate various versions of the LIWC dictionary to multiple languages over the years, allowing researchers to conduct parallel work across different languages and cultures [16]. In this study, we used the LIWC 2007 dictionary prepared in Turkish with the permission of Pennebaker. LIWC 2007 Turkish dictionary composed of 23056 words and 89 linguistic categories.

2.3. Data Cleaning

Twitter limits the length of a tweet to 140 characters. This leads users to use acronyms, remove some letters from words, and use emoticons to express special meanings. Besides a text in a tweet could contain:

1) Emoticons: facial expressions pictorially represented by punctuation and letters.

2) Hashtags: using “#” symbol to mark topics.

3) Target / mention: using the “’@” symbol to refer to other users.

The aim of the data cleaning process is to remove any unwanted content from the Twitter data. The term unwanted content is used to describe any piece of information within the tweet that will not be useful for the word count procedure to assign a class to that tweet.

The unwanted content is tabulated in Figure2.

Figure 2 Unwanted Content

Unwanted Content	ACTION
Punctuation (! ? , . ” : ;)	Removed
#	Removed
RT	Removed
Emoticons	Removed
Uppercase characters	Lowercase all content
URLs and web links	Removed
Turkish characters(ç,ş,ğ ...)	Changed(c,s,g....)
The letters after the quotation mark	Turkiye

A R script was created to read in each tweet from the database and perform processing on them in order to clean the undesirable data. This R script also served to remove stopwords from the tweets. Stop words are words such as ‘gibi (such as), belki (perhaps), which, is they have little value for sentiment analysis. Removing such words allows to use more specific words to be matched with the LIWC dictionary and hugely reduces processing during the matching stages. Examples of text before and after the data cleaning process is provided in Figure 3.

Figure 3 Pre-Post Processing Sample Table

Tweet	Status
RT @doganburak29 Unutma.... https://t.co/a0Wjau1gw8	unutma
RT @ugurdundarsozcu Sevgili Müjdat Gezen'le birlikte 144 yıllık çınar Vefa Lisesi'nin proje (!) okul yapılmasını protesto eden veli ve mezunlara destek verdik. https://t.co/Mdr3Y5344s	sevgili mujdat gezen birlikte yillik cinar vefa lisesi proje okul yapilmasini protesto eden veli mezunlara destek verdik
O kadar çok ürün çeşidimiz var ki ancak bu kadar özet geçebildik! ;) https://t.co/vXWhUUAPtr https://t.co/DGWFuxqKCy	kadar cok urun cesidimiz var ancak kadar ozet gecebildik

3. PROJECT DEFINITION

3.1. Project Objectives

First objective of this project was to develop a framework for analysis of Turkish tweets for emotion classification. To achieve our goal, we used LIWC dataset which is created by Pennebaker for emotion classification and approximately 2.2 million Turkish Tweeter data. Second objective was to see if the changes in emotions over time were related to the important events in Turkey. Third and last objective was to see if a predictive study could be done with the emotion categories.

3.2. Project Scope

Our analysis doesn't take into account many factors. Firstly, our dataset doesn't really map the real public sentiment, it only considers the twitter using, Turkish speaking people. It's possible to obtain a higher correlation if the actual mood is studied.

The prediction part of the study only took into account emotions. We know that there are many factors impact on dollar exchange such as interest rate, export of goods etc. All these remain as areas of future research.

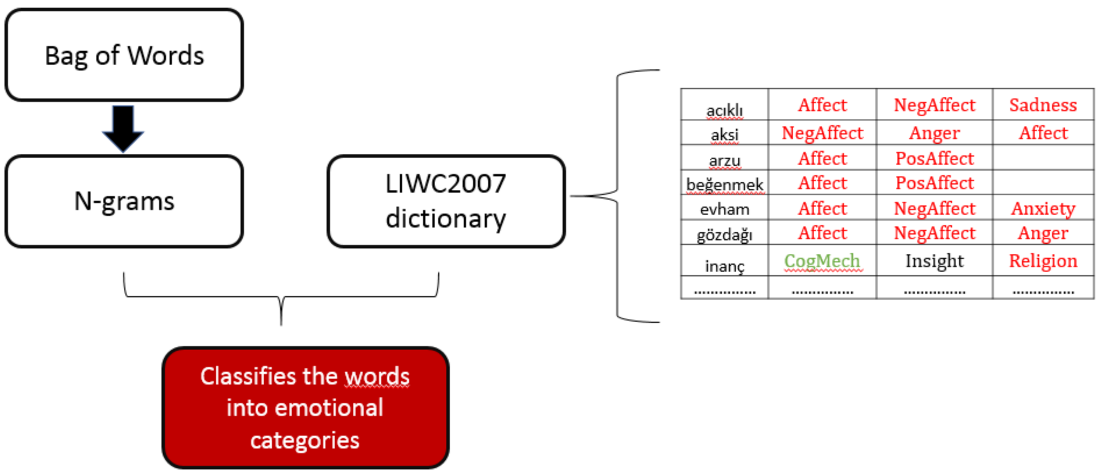
4. METHODOLOGY

We have used a two-stage approach to first create sets of words that relate to specific emotions and then using these set of emotion classifications, we applied a Genetic Algorithm (GA) and Autoregressive Model in order to predict some important events and map them with emotions.

4.1. Methods and Techniques

A two-stage approach is used in order to create emotional categories. During this process, we have used the words (N-grams) that are matched with certain emotions in LIWC and find them in the Turkish Twitter data (see Figure 4)

Figure 4 How to combine Twitter data with LIWC2007 Dictionary



We created a "bag-of-words" approach, where we first retrieved the most spoken words in the first phase of this study. Next, we extracted the sentiment of tweets using the LIWC2007 dictionary [3].

4.1.1. Bag of Words

The unit of analysis that we will use in our models will be tokens. A token is a word or group of words: 'merhaba' is a token, 'teşekkür ederim' is also a token. We can have tokens corresponding to one or two words or even series of words (n) depending on the requirements of a meaningful expression. Taking groups of n consecutive words and create a token based on that is called n-gram. In our study, we only consider 1-gram (each token is 1 word).

The Bag of Words model forms a vocabulary from all of the tweets, then models each tweet by counting the number of times each token appears in particular tweet. For example, consider the following two sentences:

Sentence 1: "The cat sat on the hat"

Sentence 2: "The dog ate the cat and the hat"

From these two sentences, our vocabulary is as follows:

{the, cat, sat, on, hat, dog, ate, and }

To get our bags of words, we count the number of times each word occurs in each sentence. In Sentence 1, "the" appears twice, and "cat", "sat", "on", and "hat" each appear once, so the feature vector for Sentence 1 is **Feature Vector 1: { 2, 1, 1, 1, 1, 0, 0, 0 },** which matches with the Vocabulary 1: {the, cat, sat, on, hat, dog, ate, and}. Similarly, the **feature vector for Sentence 2** is: { 3, 1, 0, 0, 1, 1, 1, 1}. We used ngram R package to create the bag-of-words from tokens.

4.1.2. Linguistic Inquiry and Word Count 2007 (LIWC)

To obtain a detailed understanding about the relation of words and several psychological parameters, Pennebaker and colleagues developed a computerized text analysis program named Linguistic Inquiry and Word Count (LIWC). LIWC has a dictionary composed of approximately 23000 words and 89 linguistic categories. These categories include language features (e.g., articles, pronouns), emotions (e.g. positive and negative emotion), relativity-related words (e.g., time, space, and motion) and content dimensions (e.g., occupation, death, sex, home). LIWC hierarchically organizes words based on its relevance to several categories. For instance, the word “laughed”, is categorized under several categories such as “happiness”, “positive emotion”, “overall affect,” and “past-tense verb.” [15] We ignored sarcastic, synonym and idiomatic usage of words. There have been a few recent works attempting to detect sarcasm **[22].** First, we counted all words in tweets and categorized under the specific categories LIWC2007

Turkish Dictionary **defined**[1]. Below is an example of how we operated (see Table 1 and Table 2)**.**

Sentence Example: O kadar çok ürün çeşidimiz var ki ancak bu kadar özet geçebildik!

English translation: We have so many product categories, we could just summarize this way.

Table 1. LIWC2007 Categories

1	Function	31	CanPast	61	See
2	Pronouns	32	DesireTotal	62	Hear
3	PersPron	33	DesirePast	63	Feel
4	I	34	Imperatif	64	Biological
5	We	35	Descriptive	65	Body
6	YouSing	36	Prepositions	66	Health
7	YouPl	37	Conjunctions	67	Sexual
8	HeShe	38	Quantity	68	Ingestion
9	They	39	Number	69	Relative
10	ImpPron	40	Swear	70	Motion
11	Verbs	41	Social	71	Space
12	negations	42	Family	72	Time
13	PassVerbs	43	Friends	73	Work
14	Questions	44	Human	74	Achievement
15	VerbI	45	Affect	75	Leisure
16	VerbYouSing	46	PosAffect	76	Home
17	VerbHeShe	47	NegAffect	77	Money
18	VerbWe	48	Anxiety	78	Religion
19	VerbYouPl	49	Anger	79	Death
20	VerbThey	50	Sadness	80	Assent
21	AoristTense	51	CogMech	81	Filler
22	PresentTense	52	Insight	82	Nonflu
23	PastTotal	53	Causality	83	TotalI
24	FutureTense	54	Discrepancy	84	TotalYou
25	PastDili	55	Tentative	85	TotalHeShe
26	PastMişli	56	Certainty	86	TotalWe
27	Modalities	57	Inhibition	87	TotalYouPl
28	MustTotal	58	Inclusion	88	TotalThey
29	MustPast	59	Exclusion	89	TotalMişli
30	CanTotal	60	Perception		

[1] During the data analysis, we have used R and its certain well-known libraries such as Tidyr", "Stringr" and "Dplyr" packages as well as "Tokenizers" and "timeSeries" were used".

O kadar çok ürün çeşidimiz var ki ancak bu kadar özet geçebildik!

Table 2. Distribution of words that match with LIWC categories.

O	1	2	3	8	85					
kadar	1	36								
çok	1	35	38	51	55					
ürün	74	73								
çeşidimiz	1	35	51	55						
var	1	11								
ki	1	37								
ancak	1	37	51	54	59					
bu	1	10	2							
kadar	1	36								
özet										
geçebildik	23	25	11	30	51	54	12	31	69	70

Table 3. Example of tweet classification in relation to LIWC categorization

Day1 Tweets	**word1**	**word2**	**word3**	**.......**	**word$_n$**
Tweet1	LIWC categories 10	.....	......		
Tweet2	.....	LIWC categories 2			LIWC categories 8
Tweet3	......	......	LIWC categories 15		
Tweet4	LIWC categories 10	......	LIWC categories 21		LIWC categories 8
....					
Tweet$_n$	....	....	....		....

As seen from Table 2, the word "O" is matched with categories 1 (Function), 2 (Pronouns), 3 (PersPron), 8 (HeShe) and 85 (TotalHeShe). First, for each day, we have classified each word per tweet based on distinct LIWC categories (Table 3). Next, for each day, we have aggregated the usage counts of each category in LIWC (Table 4). Finally, we calculated the percentage of each LIWC category in a given day (Table 5). For example, in day 1, for LIWC Category 1, which is about "function", there are 500 hundred words, and in total 5000 words that relate to all LIWC categories. Thus, the percentage of LIWC category 1 for day 1 is (500/5000=10%).

Table 4. Counts per LIWC categories

	LIWC categories 1	LIWC categories 2	LIWC categories 3	………..	LIWC categories 89	LIWC categories TOTAL
Day1	500	200	100		50	5000
Day2	300	150	250		150	6500
………….						
Day_n	400	500	280		440	6250

Table 5. Percentage per LIWC categories

	LIWC categories 1	LIWC categories 2	LIWC categories 3	………..	LIWC categories 89
Day1	10.0%	4.0%	2.0%		1.0%
Day2	4.6%	2.3%	3.8%		2.3%
………….					
Day_n	6.4%	8.0%	4.5%		7.0%

4.1.3. Genetic Algorithm (GA) and Autoregressive Model

In order to estimate our models, we used Genetic Algorithm (GA) based regression. The principle of the GA is that the fittest always survive and, to maintain that, part of the population will evolve to adapt better to its surroundings [23-25]. In this study "Genetic Algorithm-based" regression technique was used. The regression models that use GA generates evolve generation after generation to find better models and the best model always survives through each generation. GA begins to discover not from a point but from a pool of possible solutions. So this minimizes the risk of getting 'stuck' into a local optimum. Instead of using "trial and error" or classic statistical methods, genetic algorithms start from a model and tries hundreds of slightly different models, each model using a learning from the previous iteration, arriving at the optimum model solution in much less time than regular machine learning techniques. Genetic Algorithm is not a deterministic model so it explores all possible relations rather than assigning a-priori predictions.

According to this method, most explanatory variables are selected and the transformations to be applied to these variables are formed by Genetic Algorithm.

The prediction model is formulated with a multiple additive equation.

$$Y = \beta_0 + \theta_1 x'_1 + \theta_2 x'_2 + \theta_3 x'_3 + \cdots + \theta_n x'_n + \varepsilon$$

Y: Dependent variable, **Dollar exchange rate ($)**

x'_1 : : Explanatory variables.

The explanatory variables (LIWC categories) were included as lagged variables if the model fit was higher based on the Genetic Algorithm's choices. If the regression model includes not only the current but also the lagged (past) values of the explanatory variables (the X's), it is called a distributed-lag model. If the model includes one or more lagged values of the dependent variable among its explanatory variables, it is called an autoregressive model. Thus,

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_t$$

represents a distributed-lag model, whereas

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t$$

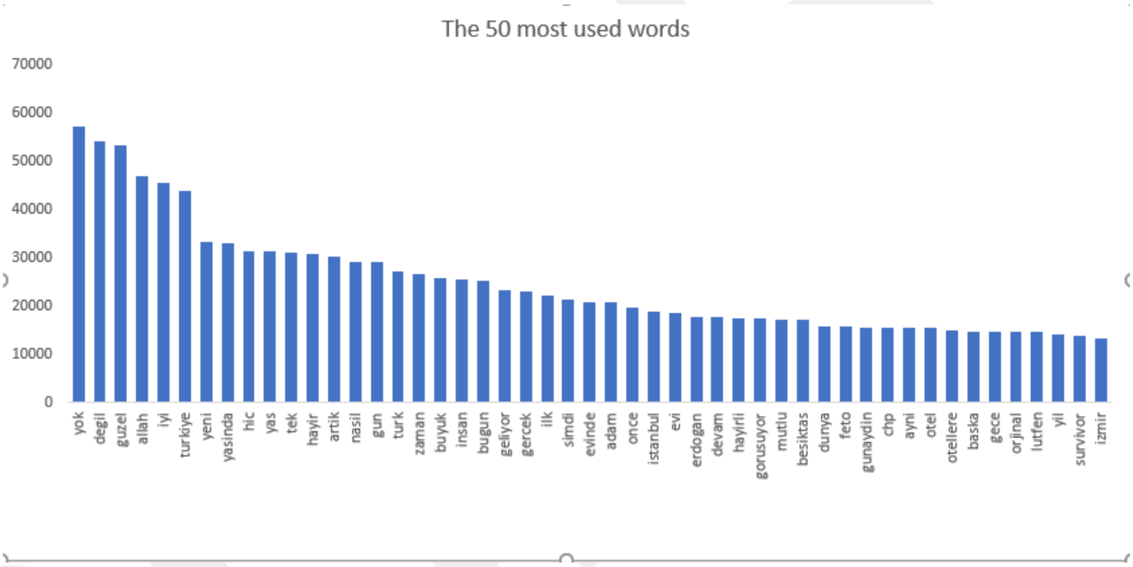
is an example of an autoregressive model. The latter is also known as dynamic models since they portray the time path of the dependent variable in relation to its past value(s) [18].

5. RESULTS

5.1. Most Used Words

Based on the results of about 2.2 million tweets in during the period of November 2016 until June 2017 (8 months), the most used 50 words are depicted in Figure 5 as well as the most used 100 words are depicted in the word cloud that is scaled based on usage frequencies (i.e., the more used ones are bigger).

Figure 5. The 50 most used words

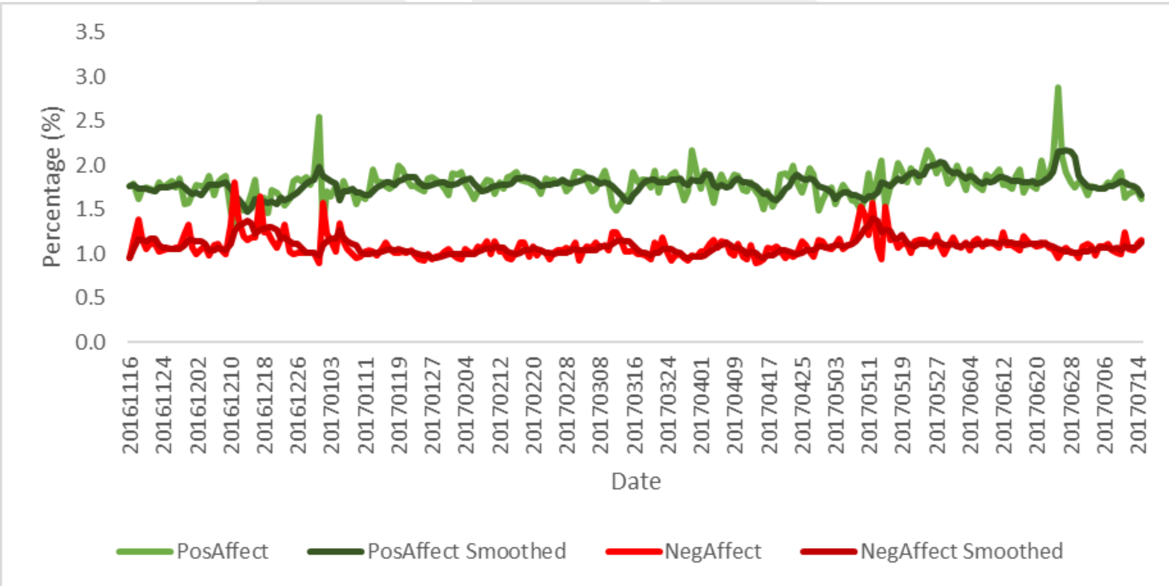


When, the LIWC categories are visualized based on usage frequency, it observed that words that are related to function (such as it, to, no, very), affect (such as happy, cried, abandon), relative (such as area, size, exit) and descriptive (such as brutal, nervous, bare) are used the most.

5.2. Emotion Tracking in Relation to Prominent Events

We tracked how emotions evolve over time based on the prominent events in and or related to Turkey. In order to make a fair comparison of the categories used over time, we used percentage values for each LIWC categories divided by the number of all LIWC categories per day instead of their nominal usage counts. Also, five-day moving averages are used to reduce the noises in the series and to see the signals more clearly. Samples of the obtained results are shown in the following figures (see Figure 6).

Figure 6 Time Series for Positive and Negative Affect



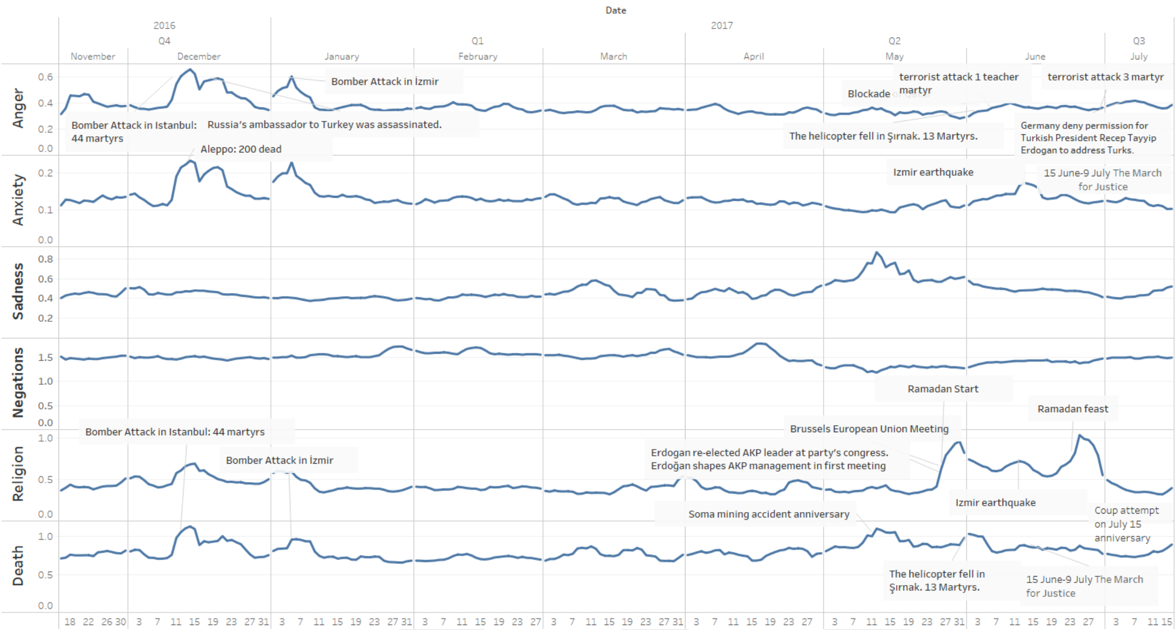
First, we track the usage of positive emotions (PosAffect) and negative emotions (NegAffect). We observe that positive emotions are expressed more compared to negative emotions. This is consistent with the word of mouth literature which shows that individuals are more likely to share positive content compared to negative content (Berger and Schwarz, 2012). Also, as expected the usage of positive and negative emotions was negatively correlated (-0.50), -0.43 for lagged values). One should realize that coding emotions as negative and positive does not fully show us the distribution over time, as it might be confounded with several distinct emotions that share the same valence but act opposing directions in a given time. For example, sadness and anger are both negative emotions, but they might be acting opposite in a given day, based on the national mood. Therefore, it is important to track diverse emotions, but not use them aggregate.

Next, we explored the relationship between a diverse set of categories including specific emotions (anger, anxiety, religion, death, sadness, achievement, leisure, desire) and their potential relation with the prominent events in the selected period in Turkey. (see Figure 7a.) As seen in the graph, especially during Dec 2016-Jan 2017 when both bombings and terrorist attacks occur there are spikes in anger, anxiety, as well as usage of words that relate to death and religion. After these periods, all categories gradually return to the baseline level approximately in 10 days. During natural disasters such as İzmir earthquake, words that are related to anxiety are used more. Further, during this period, words that are related to religious values increase. These findings are consistent with existing literature where individuals seek help from religion and spiritual beliefs to cope with difficult events. With this study, we provide field evidence to the phenomena that individuals cope with natural disasters and their related with feelings such as anxiousness with religion. As one would expect, during the Ramadan period, words that are related to religion is rapidly increasing especially at the beginning and end of the Ramadan. Further, words that are related to death show spikes especially during terrorist attacks such as the bomber attack in Istanbul, Izmir and right before the Coup attempt on July 15 anniversary.

When we explored the usage of negation over time, we notice that there is a declining pattern in the usage of “negation”, right after a peak of negation usage during the Presential Referandum. One could speculate that the usage of negation (e.g., “almiyor”, “başlamam, “değiştirmem”, “düşünmem”) could be triggered by the elections, or it might be a seasonal effect where it coincides with the beginning of the spring period, where individuals might start to be overall more positive. The anxiety emotion which has increased for a long time before the “March of Justice”, has been tending to downward trend during “March of Justice”.

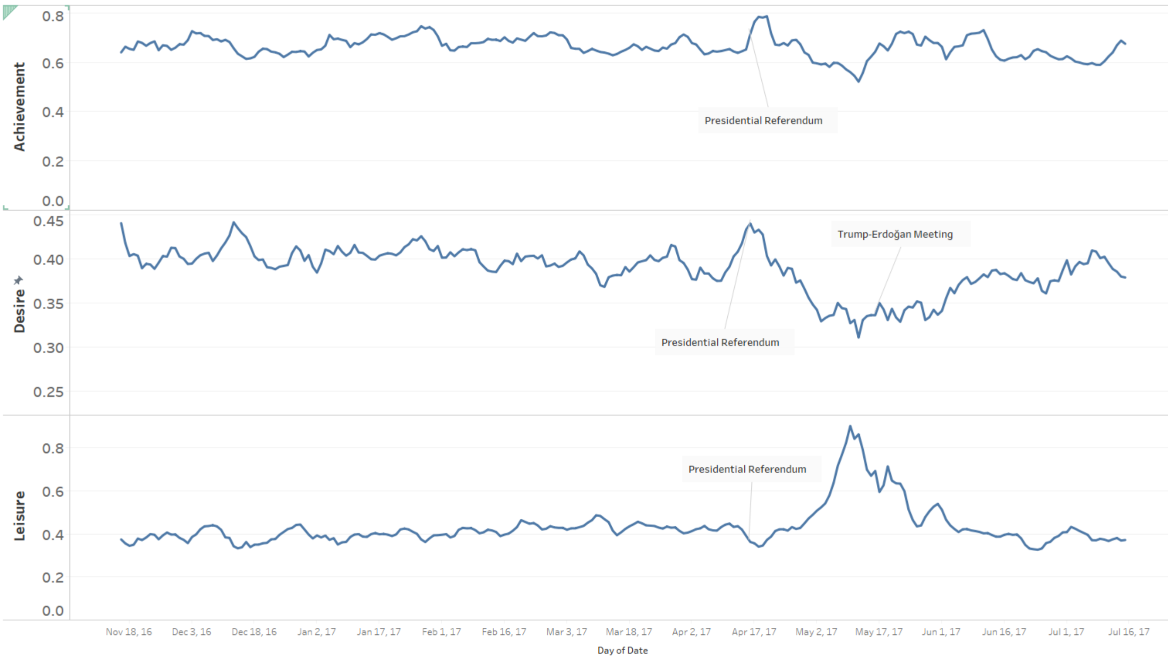
Figure. 7. Tracking public mood states from tweets posted between November 2016 to July 2017 shows public responses to some important events in Turkey

A)



Further, we explored the usage of achievement, desire and leisure (see Figure 7b). First, there is a slight increase in the usage of achievement that corresponds to Presidential Referendum. Part of the reason could be that individuals that were in favor of the referendum outcome, might have shown their feelings of achievement in their Twitter posts. The referendum seems to change the relation between the usage of words related to desire and leisure. While, the correlation between desire and leisure is -.41 until the referendum, it shows a significant leap after the day of the referendum. For instance, right after the referendum (during the period 17th April 2017 and 13th May 2017), there is a tendency towards using more words that are related to leisure (e.g., AVM, bahçe, bira, bisiklet, çakırkeyif) which at the same time, is negatively correlated (r = -.81) with using words related to desire (e.g.,). Further, the relationship between using words related to desire and leisure flips around 13th May 2017 (right after the meeting of the president of USA, Donald Trump and the president of Turkish Republic, Recep Tayyip Erdoğan). After the meeting of Trump- Erdoğan, there semms to be decline in the usage of words related to leasure, which is negatively correlated (r = -.79) with an increasing pattern in the usage of words related to desire. Finally, one of the reasons that led the usage of leisure words to increase could be TEOG exam that took place on 26-27th of April.

Figure 7B.



5.3. Dollar Exchange Rate Prediction

We know from psychological research that emotions, in addition to information, play a significant role in human decision-making. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [19]. In this research, we attempt to predict fluctuations in the dollar exchange rate based on the emotions used by public on Twitter. In our model, we used Dollar exchange rate ($) as a dependent variable, and LIWC categories that are related to emotions (Negations, desire, swear, anxiety, sadness, achievement, leisure, religion, death) as explanatory variables. We used the last 60 days as a hold-out period to avoid overfit.

In order to understand the relationship between certain emotions (anxiety, sadness and anger) and concepts (death, negation, religion, swear, leisure, and desire), we have run correlations with the dollar exchange (see Figure 8). While negation and achievement has a negative relationship, anger, death and religion has a positive relationship with the dollar exchange. This correlational evidence does not provide any causal evidence, therefore in the next step we present a forecast model to understand whether these components can predict the dollar exchange.

Figure 8.

	Dollar	Negations	Achievement	Desire	Leisure	Anxiety	Swear	Sadness	Anger	Death	Religion
Dollar	1	.477**	.272**	.213**	-.012	-.113	-.140*	-.225**	-.257**	-.380**	-.403**
Negations	.477**	1	.476**	.821**	-.607**	.182**	.246**	-.754**	.096	-.666**	-.339**
Achievem	.272**	.476**	1	.370**	-.357**	.043	.164*	-.400**	-.107	-.374**	-.069
Desire	.213**	.821**	.370**	1	-.717**	.429**	.425**	-.810**	.381**	-.493**	-.194**
Leisure	-.012	-.607**	-.357**	-.717**	1	-.485**	-.278**	.799**	-.362**	.384**	-.205**
Anxiety	-.113	.182**	.043	.429**	-.485**	1	.329**	-.346**	.854**	.297**	.325**
Swear	-.140*	.246**	.164*	.425**	-.278**	.329**	1	-.400**	.468**	-.145*	-.020
Sadness	-.225**	-.754**	-.400**	-.810**	.799**	-.346**	-.400**	1	-.250**	.645**	.066
Anger	-.257**	.096	-.107	.381**	-.362**	.854**	.468**	-.250**	1	.348**	.203**
Death	-.380**	-.666**	-.374**	-.493**	.384**	.297**	-.145*	.645**	.348**	1	.408**
Religion	-.403**	-.339**	-.069	-.194**	-.205**	.325**	-.020	.066	.203**	.408**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

When we consider the LIWC categories to be a predictor of the dollar exchange rate, the mean absolute percentage error (MAPE) of this prediction is 0.55%. The MAPE is a measure of forecast accuracy and has been widely used to compare the accuracy of models. According to the model results, R^2 is 94%. R^2 values represent the proportion of variation which is explained by the model. Hold-out period R^2 is 65%. Our results show that changes in emotions such as anxiety can be predictive of the dollar exchange rate (Figure 9). Although, the predictive model show that desire and death are predicting dollar exchange negatively, these results should be interpreted with caution as these results might mean causality between these variables. In order to test any potential causality between the predictors and the dependent variable, we have run Granger causality test using datasets with lags of different time periods shown in Eg.1 for the period of time between November 16 to July 15.

Equation 1.

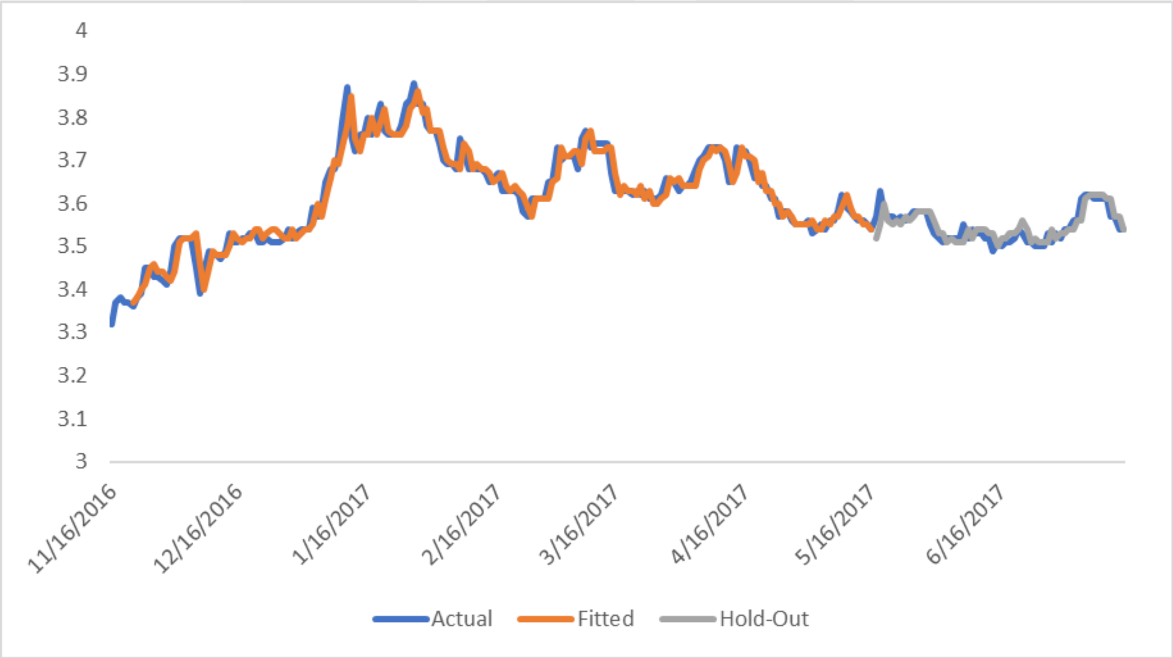
The null hypothesis is that does not Granger-cause in the first regression and that does not Granger-cause in the second regression.

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \dots + \alpha_l y_{t-l} + \beta_1 x_{t-1} + \dots + \beta_l x_{-l} + \epsilon_t$$

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \dots + \alpha_l x_{t-l} + \beta_1 y_{t-1} + \dots + \beta_l y_{-l} + u_t$$

Based on the results of our Granger causality for instance, when we use a lag up to 13 days, there is a causal relationship between anxiety and dollar exchange, which shows that a rise in anxiety increases the currency (p=0.05), yet the other way around does not hold (p=0.39). For the relationship between death and currency, we find that when the currency increases, the usage of words related to death increases.(p=0.02) While this effect might sound not intuitive, it might be explained by the effect of an exogenous variable, such as the terrorist attacks which explain the higher number of words related to death and happened to be around the times where the currency was increasing as well. Considering the effect of relationship between the usage of words about desire and currency, when currency increases, the usage of words about desire increases(p=0.04), yet not the other way around(p=0.43). It might be that when dollar increases, individuals might engage in talking about their desires and wishes (e.g., ulaşabilirse, düşmese, oynasa, konuşulmasa, inse) such as speculation of what could have been done under these economic circumstances.

Figure. 9. Predicted vs Actual Dollar Exchange Rate



6. FUTURE RESEARCH

There are some several patterns that emerged out of these research, that opens new questions. For example, it is possible to notice that religion was used more in the cases of terrorist attacks, which at the same increased the usage of death, anxiety and anger. One could argue that the usage of religion could be related to individuals seeking help from religion or coping with the negative news of death. Yet, the usage of religion might be in a negative tone, which might be derived from the religious grounds of the terrorist groups.
Further, we have raised attention the negative correlation between the usage of desire and leisure at different times, that corresponds to prominent events such as referendum and president meetings. One might wonder whether the significant flip (i.e., first leisure goes up and desire goes down, and then leisure goes down and desire goes up) during the period of the president meetings was considered negative or positive. It could be that desire might be on the rise due to expecting positive expectations in relation these two countries, or it could be that the desire might be on rise due to the negative expectations about the situation. It might be that the uncertainty makes people less likely to enjoy leisure or at least leisure in their language.

In the prediction part of the study our analysis doesn’t take into account many other external factors. We know that there are many factors impact on dollar exchange such as interest rate, export of goods etc. All these remain as areas of future research.

7.REFERENCES

[1] Gottschalk, L.A., & Gleser, G.C. (1969). The measurement of psychological states through the content analysis of verbal behavior. Berkeley: University of California Press.

[2] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: Computer Approach to Content Analysis. Cambridge, MA: MIT Press.

[3] James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The Development and Psychometric Properties of LIWC2007.

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

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

[6] Saif M Mohammad, Svetlana Kiritchenko, Using Hashtags to Capture Fine Emotion Categories from Tweets. Computational Intelligence, Volume 31, Issue 2, Pages 301-326, May 2015.

[7] Cavazos-Rehg, Krauss M, Sowles SJ, Connolly S, Rosas C, Bharadwaj M, Grucza R, Bierut LJ. (2016) An Analysis of Depression, Self-Harm, and Suicidal Ideation Content on Tumblr.

[8] Sitaram Asur, Bernardo A. Huberman (2010). Predicting the Future With Social Media.

[9] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

[10] Ghansham V.Shendge, Mangesh R.Pawar, Nikhil D.Patil, Pratik R.Pawar, Prof. Devdatta B.Bagul (2016). Real time Tweet analysis for event detection & reporting system for Earthquake.

[11] Mehmet Ulvi Şimşek, Suat Özdemir (2016). Analysis of the Relation between Turkish Twitter Messages and Stock Market Index.

[12] Z. Boynukalin. (2012). Emotion analysis of turkish texts by using machine learning methods. Middle East Technical University.

[13] Xue Zhang, Hauke Fuehres, Peter A. Gloor(2010). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"

[14] Johan Bollen,Huina Mao1,Xiao-Jun Zeng (2010). Twitter mood predicts the stock market

[15] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015

[16] Ryan L. Boyd (2017). Psychological Text Analysis in the Digital Humanities

[17] Çiğdem Köşe Demiray (2015). Change in Usage of Cognitive Words, Affect Words, Discourse Markers, and First Person Pronoun Position Among Clients in Psychotherapy Process.

[18] Damodar N. Gujarati, Dawn C. Porter (January 2009). Basic Econometrics.

[19] Nofsinger, J. (2005) Journal of Behaviour Finance. 6, 144–160.

[20] Mohammed ALSADI, Sevinç GÜLSEÇEN, Elif KARTAL (2016). TOP 10 Turkish Universities Twitter analysis user sentiment analysis and comparison with international ones.

[21] Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz. Predicting Depression via Social Media.

[22] Liebrecht, C.C. ; Kunneman, F.A. ; Bosch, A.P.J.(2013). The perfect solution for detecting sarcasm in tweets #not

[23] Nabiyev V.V. 2003. “Yapay Zeka”.

[24] Wright, A. (1991). “ Genetic Algorithms for real parameter optimization,” in G.J.E. Rawlins (Ed.), Foundations of Genetic Algorithms 2, San Mateo, CA: Morgan Kaufman

[25] Michalewicz, Z., 1994, Genetic Algorithms + Data structures = Evolution Programs, 2nd ed., New York: Springer-Verlag.

APPENDIX

The R code for data cleaning part.

```
sel_cleantext <- function(mytext) {
  require(tm)
  mytext <- gsub('ğ', 'g', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('Ğ', 'g', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('ü', 'u', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('Ü', 'u', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('ş', 's', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('Ş', 's', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('İ', 'i', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('I', 'i', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('ı', 'i', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('ö', 'o', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('Ö', 'o', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('ç', 'c', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub('Ç', 'c', mytext, ignore.case=TRUE) ## turkce karakter

  mytext <- gsub(xtr_s, 's', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub(xtr_I, 'i', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub(xtr_i, 'i', mytext, ignore.case=TRUE) ## turkce karakter
  mytext <- gsub(xtr_g, 'g', mytext, ignore.case=TRUE) ## turkce karakter

  mytext <- gsub('\n', ' ', mytext, fixed = TRUE) ## enter
  regex_http <- "?http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+"

  mytext <- gsub(regex_http, " ", mytext) ## http ile baslayan kelimeler
  mytext <- gsub("htt…", " ", mytext, ignore.case=TRUE) ## http ile baslayan kelimeler
  mytext <- gsub("http:/…", " ", mytext, ignore.case=TRUE) ## http ile baslayan kelimeler

  mytext <- gsub("<.*?>", " ", mytext) ## multiple html tags

 mytext <- gsub("<.*?>", " ", mytext) ## multiple html tags

 gsub("'\\S+\\s*", " ", mytext) ## tirnak ayrci sonrasi harfler Turkiye'nin
 mytext <- gsub("@\\S+\\s*", " ", mytext) ## @ isareti ile baslayan kelimeler
 mytext <- gsub("&gt", " ", mytext) ## > &gt isareti ile baslayan kelimeler

 mytext <- gsub("3g", "ucg", mytext, ignore.case=TRUE) ## numerik istisnalar
 mytext <- gsub("4g", "dortg", mytext, ignore.case=TRUE) ## numerik istisnalar
 mytext <- gsub("5g", "besg", mytext, ignore.case=TRUE) ## numerik istisnalar
 mytext <- gsub("4.5g", "dortbucukg", mytext, ignore.case=TRUE) ## numerik istisnalar

 mytext <- gsub("[^[:alpha:]///' ]", " ", mytext)

 mytext <- gsub("#", " ", mytext) ## hasstag gereksiz

 mytext <- gsub("\\s.\\s", " ", mytext)
 mytext <- gsub("^.\\s", " ", mytext)
 mytext <- gsub("\\s.$", " ", mytext)

 mytext <- gsub("      ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("     ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("    ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("   ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("  ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("  ", " ", mytext) ## aradaki cift bosluklar
 mytext <- gsub("(^ +)|( +$)", "", mytext)  ## bastaki ve sonraki bosluklar


 mytext <- gsub("[[:punct:]]", " ", mytext) ## removePunctuation
 mytext <- tolower(mytext)
 mytext
}
```

The R code for word counts:

```
library(tokenizers)

birlesik$Contents_en <- sel_cleantext(birlesik$Contents)
head(birlesik$Contents)
metinler <- birlesik$Contents_en

names(metinler) <- birlesik$GUID


head(metinler)



Sys.time()

  xdatselml.text <- data.frame(text = metinler, xbin = 1, stringsAsFactors = FALSE)
  xdatselml.text$xbin <- ((1:nrow(xdatselml.text)) %% (ceiling(nrow(xdatselml.text)/800))) +1
  for (i in 1:max(xdatselml.text$xbin)) {
    xdattextsplit <- subset(xdatselml.text, xbin == i)
    xdattextvec <- as.character(xdattextsplit$text)
    names(xdattextvec) <- rownames(xdattextsplit)

    xx3tmp <- melt( tokenize_ngrams(xdattextvec, n = 1) )
    if (i == 1) { xx3  <- xx3tmp } else { xx3 <- rbind(xx3, xx3tmp) }
    print(paste(i, Sys.time(), nrow(xx3), sep = " | "))
    gc()
    system(shQuote("C:/apps/cleanmem/CleanMem.exe" , type = "cmd" ))
  }
```

Genetic algorithm based regression analysis metrics.

Test Period

R^2	93.66%
Adjusted R^2	93.52%
Durbin-h	0.232728184
Degrees of Freedom	172

Mean Absolute Percentage Error: 0.55%

Independent Variable	Lag Periods	Coefficient	T-Statistic
Constant		0.334772634	3.289
Lag Dependent Variable	1	0.940177788	46.899
DesireTotal	5	-0.245238032	-1.849
Anxiety	5	0.372732835	2.996
Death	5	-0.088519402	-2.485

Hold-Out Period

Hold-Out Periods	60
Mean Absolute Percentage Error	0.43%
Hold-Out R^2 Statistic	64.48%