Turkish Text Analysis System for Automatic Detection of Psychiatric Disorders

Zeynep Orhan, Mine Mercan

Computer Engineering Department
Fatih University
İstanbul, Turkey
zorhan@fatih.edu.tr,
minemercan@fatih.edu.tr

Ahmet Sertbaş
Computer Engineering Department
İstanbul University
İstanbul, Turkey
asertbas@istanbul.edu.tr

Abstract

This study deals with a Turkish text analysis system for adults to detect psychiatric disorders. The data is organized as answers of two questions for each subject. The aim of this analysis system is to determine wether the subjects are healthy or suffering from a psychiatric disorder, namely Depression or Anxiety based on their language usage. Naïve Bayes, Support Vector Machine and Decision Tree ML methods are used. Words or categorised words are used as features. The aim of this research is presenting a psychiatric disorder automatic classification system using the ML methods as one of the initiators of Turkish natural language processing researches.

1 Introduction

In this era technology has become an essential part of our lives. Almost all of our activities depend highly on internet technology. Using computers, smart phones and tablets are far or less essential in the activities of daily life. Today people have easy access to data through technology and this leads to many changes. In today's world people prefer shopping online and use social blogs to share their comments, pictures or videos in the virtual environment. The companies gather feedbacks about their products from these customers. This fact makes the internet that is an ocean of digital data a far easier and effective resource of data analysis tools compared to the others. The data on the internet is generally in the form of text that can be obtained from comments, blogs, or any other platform where people share their opinions. Additionaly, the data on the internet also reflect a variety of cultural, locational, ethnic, etc differences. Furthermore, generally people are not under the effect of any other person when they share something on the internet, especially when they share their own opinion on social blogs, etc. this type of data is more reliable and easy to analyze compare to the other resources. Last but not the least, using internet can also be very useful for scientists who study about any statistical data about people or people's opinion.

Opinion Mining and Sentiment Analysis (OMSA) researches are increased due to the easy access to huge data on internet. OMSA studies can be used to get hints about the psychological state of the people. The use of language provides important evidence about the human mind, including emotional ups and downs, altering moods caused by psychiatric disorders [5]. The diagnosis of psychiatric disorders can be aided by computer applications. Natural Language

Processing (NLP) and Sentiment Analysis (SA) are commonly employed in the technological applications of these areas. NLP serves to design and build systems which analyze and generate human languages. NLP techniques use Machine Learning (ML) algorithms on the derived data. ML techniques have become widely applicable in many different areas due to technological developments and recent internet opportunities.

SA refers to the use of NLP, text analysis and computational linguistics to identify and extract subjective information in source materials. SA is the automatic classification of a text, trying to determine the attitude of the writer with respect to a specific topic. The attitude may be the judgment or evaluation, the feelings or the intended emotional communication [12]. Text categorization according to affective relevance, opinion exploration for market analysis, etc., are examples of applications of these techniques. Opinion Mining (OM) is the process of automatic extraction of knowledge from the opinion of others about some particular topic or problem [2].

Moreover, OMSA can be used to identify the psychological state of a person by the help of the computer applications. Psychiatric disorders are one of the great challenging problems of modern world. The wide spread effects of this fact are influencing deeply the humanity and are particular concern to not only the individuals, but also their family members and to the whole society. Nowadays, most people suffer from psychiatric disorders. Psychiatric disorders have high risks and impacts. Early detection of these disorders could be helpful to get better and quick results. Majority of the frequently encountered disorders are often under-diagnosed and under-treated [3]. Failure to intervene early and effectively impacts individuals and their family members adversely and results in profound long-term costs to society. The standard approach to diagnosing psychological health disorders is through a series of clinically administered diagnostic interviews and tests [4]. However, assessment of patients using these tests is expensive and time-consuming. Furthermore, the stigma associated with mental illnesses motivates inaccurate self-reporting by affected individuals and their family members, thus making the tests unreliable [5]. There is a growing interest in the area of developing computer applications that analyze psychiatric disorders day by day due to the aforementioned reasons.

This paper demonstrates a text analysis system to detect psychiatric disorders from the texts written by them. The rest of it is structured as follows. Section 2 defines the methods and describes the data set. Section 3 describes the system with proposed features and classification results. Finally, Section 4 concludes with a description of the impact of this work.

2 Methodology

2.1 Data

The data was obtained from a psychiatry clinic, namely NPİstanbul Neuropsychiatry Clinic. There were 267 subjects included in the study: Part of them is diagnosed with psychiatric disorders, such as depression (DPR), anxiety (ANX) and part of them are used as healthy controls (HC). Their demographic information was gathered such as gender, age, etc.

In order to collect texts from subjects, two questions were asked to all of the applicants and they wrote their answers down. Subjects were asked their thoughts about life, future and daily activities. The personal information that depicts the identities is removed for privacy concerns. The responses are believed to provide indications of attitudes, beliefs, motivations, or other mental states.

The data is separated with "/" sign. It starts with personal information about a subject's given identity number, gender and age. The next three fields include information related to the diagnosis given by the doctors, such as depression, anxiety and normal. Last two fields are the responses of the subjects for the questions mentioned before. An example of one subject's data is given in Figure 1.

267 subjects took part in the study. According to this data, 59 subjects were diagnosed with DPR, 64 with ANX and other 49 were diagnosed as HC controls. Some of the subjects were excluded from disorder classification due to sparseness of other types of diagnosis.

Finally the total subjects that are included in the study are 108. Distribution of subjects according to gender and disorders are provided in Table 1.

Id / Gender / Age / Education / Diagnosis1 / Diagnosis2 / Diagnosis3 / Response1 / Response2
4/K/20/lise/depresyon/ / /Kendim hayata genel olarak karamsar bakan ama sürekli hayatında mutluluğu arayan birisi olarak açıklanabilir. Kendimle ilgili yeni şeyleri öğrenmek istiyorum Fakat isteğimi eyleme geçirmede zorluk yaşıyorum ve bu doğal olarak hayatımı etkiliyor. Ben kendisiyle barış içinde olan değil hatalarından, yanlışlarından dolayı sürekli kızan biriyim bu yüzden de kendimle sürekli bir savaş halindeyim. Bu da hayatımı negatif anlamda etkiliyor./Bugün dışında herhangi bir gün. Sabah kalktım kahvaltımı yaptım. Bilgisayarda film izledim. Öğle yemeğimi yedim, televizyon seyrettim. Arkadaşlarımla sinemaya gittim. Eve geldim tekrar bilgisayarda film seyrettim. Yatmadan önce bir saat kitap okuyup uyudum./

Figure-1.Text data of one subject.

Table-**1.** *Distribution of subjects according to gender and disorders.*

Diagnosis	Gender		Total
	Male	Female	
DPR	26	33	59
ANX	26	38	64
HC	15	34	49

2.2 Methods

The language analysis of texts is carried out in order to determine whether the subjects are healthy or suffer from any psychiatric disorder.

The Harvard-III Psychological Dictionary [6] (Turkish version) was used to find

categories of words. A Turkish version of the Harvard III dictionary was generated by translating the original Harvard-III Psychological Dictionary. In the translated version of this dictionary, 4500 words were grouped into 83 categories. It included social, cultural and natural objects, behavioral and psychological processes, qualifiers, institutional contexts, status connotations and psychological themes [6].

In this study Turkish morphological analyzer and disambiguation tools [7] and Weka [8] programs were used. During the pre-processing of the collected data by means of morphological analyzer, words are converted into root-affix combinations with parts of speech such as Noun, Adj (adjective), Adv (adverb), Verb, Pnon (pronoun), etc. Each sentence is written between <S><S> and </S></S> tags. Disambiguator is a natural language processing application that tries to determine the intended meaning of a word or phrase by examining the linguistic context in which it is used. Each line of the morphological analysis result refers to a word, which is composed of the original word, its root and its morphological analysis results. Since there could be more than one morphological analysis result for a typical Turkish word, a disambiguator tool was required. After disambiguation process, the most proper result is chosen by the tool. Table 2 shows an example of a morphological analysis result for an ambiguous word. First three rows are the results, the last row is the selected one. Figure 2 shows an example sentence as the output of the morphological analyzer and the disambiguator.

Table-2. An example of a morphological analysis result for an ambiguous word with the selected one.

Original	Root	Morphologic analysis

word		
olacağım	ol	+Verb+Pos+Fut+A1sg
olacağım	ol	+Verb+Pos^DB+Adj+FutPart+P1sg
olacağım	ol	+Verb+Pos^DB+Noun+FutPart+A3sg+P1sg+Nom
olacağım	**ol**	**+Verb+Pos+Fut+A1sg**

```
<S> <S>
başarılı
başarı+Noun+A3sg+Pnon+Nom^DB+Adj+With
olacağım ol+Verb+Pos+Fut+A1sg
. .+Punc
</S> </S>
```

Figure-2. Disambiguation of morphological analysis result of a sentence.

3   Experiments

3.1 Experimental Setup and Result

The data was grouped under three categories: Depression, Anxiety and HC. The first two diagnosis group and HC group was classified into two classes as patients and healthy persons. Among each group, the major part of the data was used for training and a small amount was reserved to test the system.

The features used in the experiments were extracted by the content analysis of the textual data. The words from all the subjects' texts were analyzed using the morphological analyzer then the morphological disambiguator program, which was used to determine the roots of words. Features were extracted out of the roots of the words obtained from the training data.

Two methods were applied. The method uses the binary information of the existence or nonexistence of a word in the prepared

set of features for each subject (Table 3), while the second one provided also information about the existence or non existence the usage of each word category (Table 4). The Harvard-III Psychological Dictionary [6] (Turkish version) was used to find categories of words.

Table-3.Training data including all distinct words as columns and binary values indicating word existence in subject text.

Id	Gender	W_1	W_2	W_3	W_4	...	W_n	Class HC (N:Neg P:Pos)
125	M	0	1	0	0	...	1	N
41	F	1	0	0	1	...	1	P
134	M	1	0	0	1	...	0	P

Table-4. Training data including all distinct category as columns indicatiing category existance in subject text.

Id	Gender	C_1	C_2	C_3	C_4	...	C_n	Class HC (N:Neg P:Pos)
125	M	1	1	0	0	...	1	N
41	F	1	0	0	1	...	0	P
134	M	0	0	1	1	...	1	P

3.2 Classification

Classification is achieved using ML methods provided by a data mining tool named Weka [9]. Weka provides many well-known ML algorithms as classifiers that can be applied to various data. It has the utility to test and train the classifiers. It provides results of each training and testing process. In this study for the classification of subjects, Naïve Bayes (NB), Sequential Minimal Optimization (SMO), and a decision tree algorithm J48 is used.

NB is a Weka class for a NB classifier using estimator classes. The classifier is not an updateable classifier, because the estimator precision values are chosen based on the analysis of the training data [9].

SMO algorithm of Weka trains a Support Vector Classifier (SVC) with John Platt's SMO. A SVC is a classifier which takes a set of data and for possible two classes predicts the membership of each data according to those classes [10].

Table-5. The distribution of number of subjects that were used for training and testing experiments.

	HC	DPR	HC	ANX
Training	39	49	39	54
Testing	10	10	10	10

Table-6. The accuracy of the systems for the DEPRESSION and control data with binary values.

Method	Words	Categories
NB	80	60
SMO	70	60
J48	60	65

J48 is a decision tree algorithm. In Weka it is a class for generating a pruned or unpruned C4.5 decision tree [11].

The features in Table 3 and Table 4 are used to train the system with the aforementioned classifiers. The distribution of the data for each group is provided in Table 5. The test results for each of these classifiers of the data with binary values are given in Table 6 and Table 7.

Table-7. The accuracy of the systems for the ANXIETY and control data with binary values.

Method	Words	Categories
NB	60	50

SMO	60	60
J48	50	70

4 Conclusion

This research contributes in detecting psychiatric disorders of adults by proposing a text analysis system. Since the new generation is expressing themselves mostly in electronic environment, the study gains an important impact in detecting psychiatric disorders in people. There are a few studies using NLP methods in Turkish text analysis.The study examined three groups depression, anxiety, and healthy, which were pre-diagnosis by psychiatrics.

This study presents psychiatric disorder automatic classification using the ML methods as one of the initiators of this field. The distinct words and their categories used by the subjects were used as features for the ML techniques NB, SMO, and J48 of Weka library. The promising results of the study show that language usage can be good indicator of psychological state of the students. It is an important research, since there are few studies using NLP methods in Turkish psychological text analysis.

In the future, it is planned to obtain the semantic ontologies and dictionaries that can be used in these types of researches and specific to various disorders. Additionally, it can lead to practical economical tools that can be used to early diagnose the disorders of people.

5 References

[1] EROĞUL, U. (2009). Sentiment Analysis in Turkish, Master's thesis, Middle East Technical University, Ankara Turkey.

[2] SINDHU, R, RAVENDRA R, SINGH J, RAKESH R. K., "A Novel Approach for Sentiment Analysis and Opinion Mining.", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 4, 2014

[3] KESSLER, R. C., ZHAO, S., KATZ, S.J., KOUZIS, A.C., FRANK , R.G., EDLUND, M,. Leaf, P. (1999). Past-year use of outpatient services for psychiatric problems in the National Comorbidity Survey. Am J Psychiatry, 156115- 123

[4] WEATHERS, F. W., KEANE, T. M., & DAVİDSON, J. R. T. (2001). Clinician-Administered PTSD Scale: A review of the first ten years of research. Depression and Anxiety, Vol 13(3), 132-156.

[5] SALEEM, S., PRASAD, R., VITALADEVUNI, S. N. P., PACULA, M., CRYSTAL, M., MARX, B., SLOAN, D., VASTERLING, J. AND SPEROFF, T. (2012). Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts. In COLING (pp. 2375-2388

[6] JAMES C. MONTAGUE, JR., The Effect Of Institutionalization On The Social Behavior And Language of Mentally Retarded Children, University of Florida, (1971), pp: 93-109.

[7] OFLAZER. K, 1993. Two-level description of Turkish morphology. In Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, April. A full version appears in Literary and Linguistic Computing, Vol.9 No. 2, 1994.

[8] WITTEN, I. H. AND FRANK, E., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2 edition, 2005.

[9] John, G. H., AND LANGLEY, P., "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1996, pp 338-345.

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

[11] QUINLAN, R., (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[12] KAUR, A., GUPTA, V., (2013). A Survey on Sentiment Analysis and Opinion Mining Techniques, Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 4, November 2013.