Sentiment Analysis And Opinion Mining For Turkish Tweets

Mehmet ÖZGEN

Bil-711 Project Report

N14223406

***Abstract-* This paper gives information about how to do analysis on Turkish Tweets, Weka application information and Turkish Sentiment Analyzer Demo application.**

I. INTRODUCTION

This paper considers the problem of classifying Turkish Sentences and determining whether they are positive or negative. And I use three machine-learning methods (Naive Bayes, Decision Tree, and Support Vector Machines) for testing my analysis with the help of Weka API.

*Twitter Application in Turkey*

Nowadays, Internet is a sine qua none for our life. What happens to us we write on the Internet via social networks like twitter.

On the other hand opinions are central to almost all human activities because they are key effecters of our behaviors [1]. Whenever we need to make a decision, we want to know others’ opinions. Because we always need our friend’s opinion. In real world, businesses and organizations always need information about what their customer’s wants and public opinions about their products. This paper gives some information about how to get public opinion about the product with the help of social media, Twitter.

Twitter is getting more and more popular day by day in Turkey. There are 12 Million people use Twitter in Turkey and almost 7 Million Tweets are written in a day. So this is an amazing database for data miners. An addition to this twitter API is the easiest tool for developers.

On the other hand, people are forced to write simple in order to fill 140 characters effectively. People just write what they think. This is core opinion of the writer. However sometimes this is may not be so simple, %47 percentages of people whose age rages are 18-29 use twitter and these people tend to use abbreviations while writing a tweet. For instance: ‘iim’ instead of ‘iyiyim.’Because of this situation we encounter so many ambiguities in the analysis. Opinion Spamming also is a major issue. Fake accounts users promote or discredit target products and services without disclosing their true intentions. This paper focuses on the Turkish tweets that have not any abbreviations. Thus, it is often easier to archive high sentiment analysis accuracy.

In most applications, one needs to analyze opinions from a large number of people. This indicates that some form of summary opinions is desired [1]. Fig.1 is used in this demo.

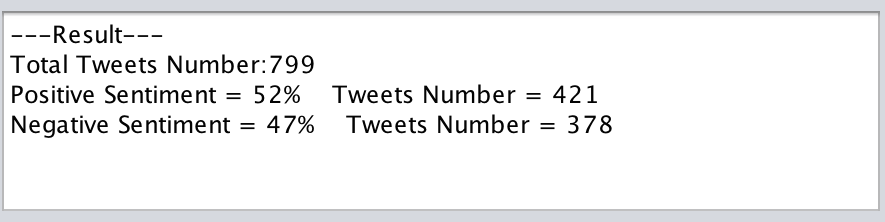


Fig. 1. Result

II. SENTIMENT ANALYSIS RESEARCH

In general, sentiment analysis has been investigated mainly three levels:

1. *Document Level*

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment [2]. We focus on the review of a product and express overall positive or negative opinion about the product.

1. *Sentence Level*

The task at this level goes to the sentences and determiners whether each sentence expressed positive, negative or neutral opinion. Neutral usually means no opinion. This level analysis is closely related to subjectivity classification, which also distinguishes the sentences. The analysis in this paper is related to the Sentence Level. In the demo, we analyze the Turkish sentences positive (Olumlu) or negative (Olumsuz).

1. *Entity and Aspect Level*

Both of the levels, which are discussed above, do not discover what exactly people liked and did not like. Aspect level directly looks at the opinion itself. It is based o the idea that an opinion consists of a sentiment (Negative or Positive) and a target of opinion [1]. That means opinion targets are described by entities. For example: “IPhone’s battery life is too short.” Entity is IPhone.

As Turkish is an agglutinative language, Sentiment analysis on it is more complicated. A word can get more than two derivational affixes but the word structure cannot be changed. This makes morphological analysis more significant.

Researchers have proposed many approaches to compile with sentiment word. Three main approaches are: manual approach, dictionary-based approach and corpus-based approach. Turkish has not got any particular dictionary for sentiment analysis. Corpus-based approach is more suitable than the others. Also the word’s meaning sometimes depends on the context. For Example: “Phone speaker is quiet.” Quiet has got negative meaning in the sentence; “the car is so quiet” at this time quiet has got positive meaning. Corpus-based Approach, Hatzivassiloglou and mckeown [3] propose a novel approach to extract semantic orientation of a set of adjectives based on linguistic features. They try to find semantic orientation of other adjectives in large corpus with using a set of linguistic constraints and a small list of opinion adjective words. They look whether an adjective is linked to another one known as positive or negative by a conjunction or disjunction or not. They use four steps in this method:

1. All conjunctions of adjectives are extracted from the corpus.

2. Using a log-linear regression model constructed with training set, it is determined that whether each two conjoined adjectives are same or opposite orientation. According to the result of the model on the different test sets, a graph is obtained. Nodes of it are adjectives and links of it show orientation between adjectives.

3. According to a clustering algorithm, the graphs are partitioned into two sets.

4. Items are labeled as positive and negative based on frequency of positive adjectives in the sets.

Corpus-based approach is not effective than dictionary, but it is more practical and agile. However, in this demo we try to do our corpus with 799 tweets positive or negative meaning with the help of zemberek that is java-based morphological analyzer.

III. EXPERIMENTAL TESTS

*A. Weka*

Weka is one of the machine learning packages. It is developed on Java Platform in Waikato University. It has got GPL (It was developed by Richard Stallman in 1983.) license, which supply facilities to users and developers. Its explicit name is Waikato Environment for Knowledge Analysis.

Weka reads data from simple formatted file. It recognizes the arbitrary changes from data and it accepts them as a numerical and nominal values. Also it attracts data from database, but data format must be convenient to the file format. Arff, CV5 and C4.5 are kinds of file format for Weka. In this demo, I prepared my data file as a Arff file format and saved the document with the extension of txt. Arff (Attribute Relation File Format) is a kind of ascii text file which contains data and attributes. In this demo, I trained Arff file with three methods, which are decision tree, naive Bayes and support vector machine.

Kappa statistic calculating is an alternative way for evaluating correction of the data. Kappa Statistic is a general term, used in categorized data. Typically, Kappa is used for assessing the degree of acceptance two or more observers with the help of examine these categories (1).

*K = (Po-Pc)/(1-Pc) (1) (Po: accepted proportion, Pc: predicted to accept proportion)*

Support Vector Machine (SVM) is a kind of machine learning method, which is used for classification of the data. Drawing a plane between two groups of data is a good way for classification. And SVM decides on how to draw the plane of classification via this formula (2).

*(2)*

Naïve Bayes is an algorithm that identifies which class the data belongs. It uses probability science for classifications via the formula (3).

 *(3)*

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie et al., because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate [4].

*B. Zemberek*

Zemberek [5] is an open source code Turkish Language Processing library and OpenOffice, LibreOffice. Its first sale is distributed with BSD license. It is developed by Java and it has orthographic control, suggestions for false words, spelling, deascifier[6], false coding cleaning functions. For Second sale, which is Zemberek 2 MPL license is used for setting a DDI substructure for all Turkish languages, all architectural changes are applied. A server, which is constructed by using Zemberek, supports Pardus in orthographical controlling. Server communicates with other applications on TCP-IP sockets with a protocol, which is similar to ISpell. Because of being developed in Java, it is platform free. In this demo I used Zemberek 2.1 version. In this demo, we use Zemberek in order to split stems in the tweets.

*C. Experiment*

In this demo,

1. First run the MainGui.class.
2. Choose file path of data, positive and negative txt file and write output file name.
3. Push the result button in order to create output file.
4. Choose attribute file path and write arff file name.
5. Push the create arff file button.
6. Select the training algorithm.
7. Push the train button and see the result.

Initially we use 4 file which are data, positive, negative and attributes files. Attributes file includes all positive and negative file data. Data file set of data that contains 799 row tweets. When we push the result button, the algorithm creates 3 files, output.txt, which is analyzed data, output.txt.zem, which is analyzed data with Zemberek, output.txt.tmp, which is sentimentally analyzed. And then we create arff file in order to use Weka libraries. Choose training algorithm, after pushing result button results are seen like Fig.2, Fig.3, and Fig.4.

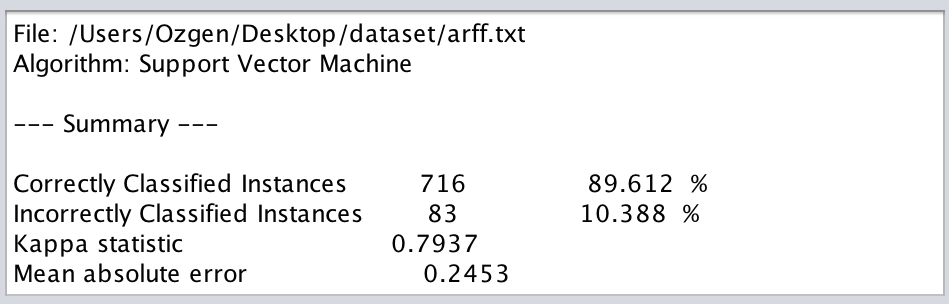


Fig. 2. SVM Result

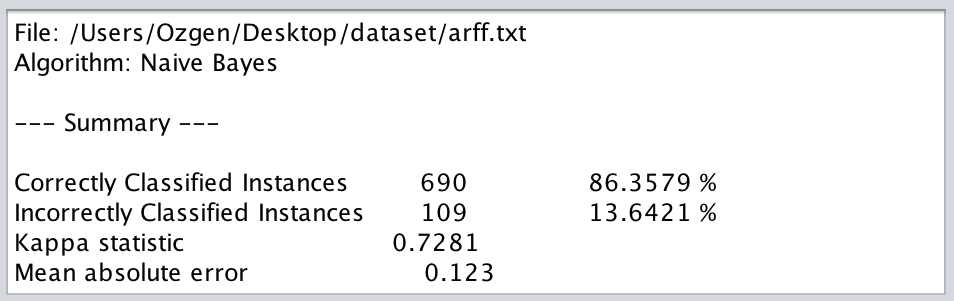


Fig. 3. Naïve Bayes Result

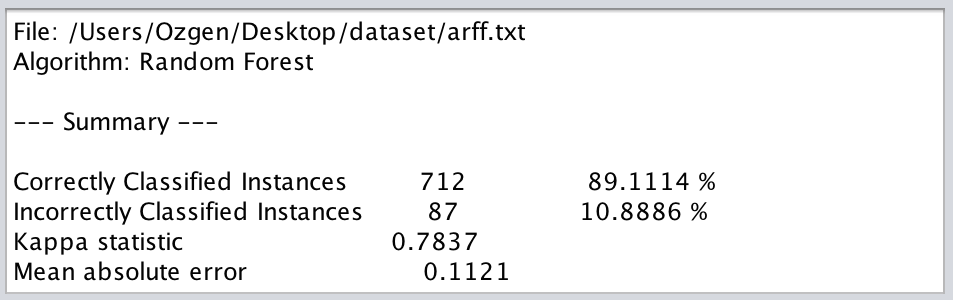


Fig. 4. Random Forest Result

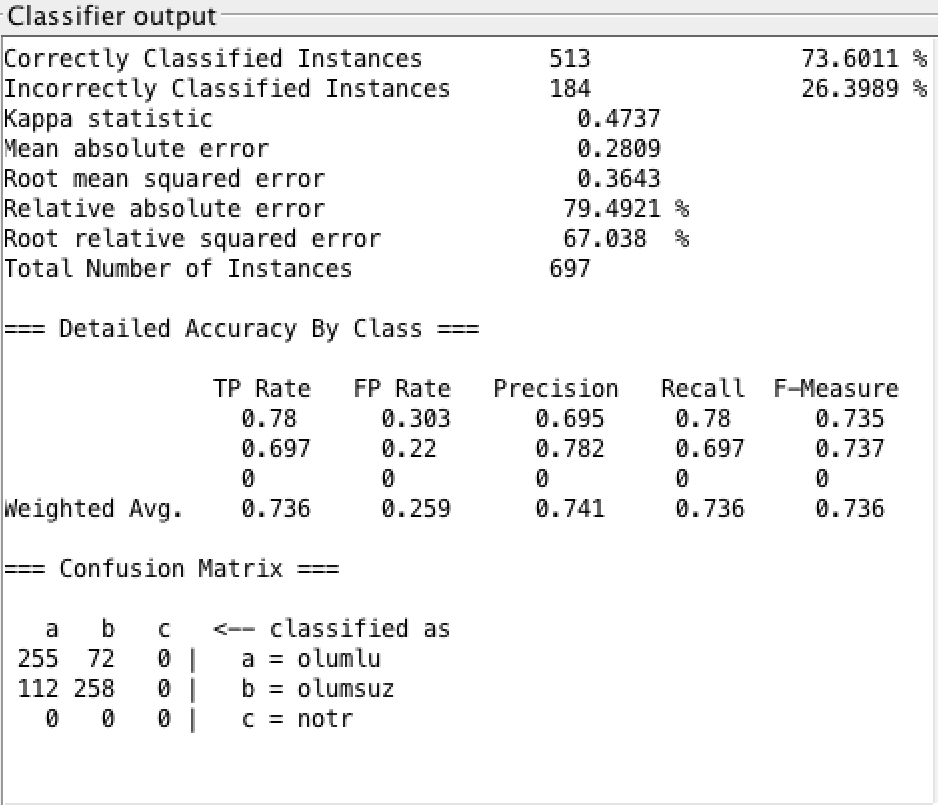


Fig. 5. SVM Result in new Data (697 tweets) with Weka GUI

After the training of the 799 tweets in our corpus, I created new dataset, which has 697 tweets. When we look at the result in new Data via Weka GUI, the dataset is classified %73 correctly with the help of the corpus we created. (Fig.5.) This result is not satisfied with several reasons that explain next title.

IV. RESULTS

I analyzed 799 Turkish Tweet with two categories, positive or negative. We do 3 kinds of train, and then I create new dataset (697 tweets). I use Weka GUI for getting prediction results. The result of the training shows that average %73 of the data is correctly classified. And Support Vector Machine method gives the best result in classification rate. However, the analysis is not satisfied with these reasons:

1. Data is analyzed without any corpus.
2. Attributes are not extensive enough.
3. The analysis depends on the negative or positive words these may not be related with the context meaning.
4. The algorithm, which has positive and negative counter, decides the sentence negative or positive but some words have effective meaning than the others, but we ignore this situation.
5. “The data we analyzed is true or not?”, “How to find fake tweets?” , these are not answered with this study.

But this study, as I use open source library, is open to be developed and after still has the problems above, the demo is hope promising for the future.

REFERENCES

[1] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167.

[2] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[3] M. K. Hatzivassiloglou, V. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, New Brunswick, NJ. P. 174-181

[4] Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: springer, 2009.

[5] <http://tr.wikipedia.org/wiki/Zemberek>

[6] Akın, Ahmet Afsin, and Mehmet Dündar Akın. "Zemberek, an open source nlp framework for turkic languages." Structure 10 (2007).