An NLP STUDY for PUBLIC TRANSPORTATİON

İsmail Yavuz
Department of Computer Engineering
Marmara University
Istanbul/Turkey

Arif Taha Ay
Department of Computer Engineering
Marmara University
Istanbul/Turkey

Faruk Furkan Şişman
Department of Computer Engineering
Marmara University
Istanbul/Turkey

Oğuzhan Demirhan
Department of Industrial Engineering
Marmara University
Istanbul/Turkey

Nur Kızıldağ
Department of Industrial Engineering
Marmara University
Istanbul/Turkey

Habib Burak Can Erdönmez
Department of Bio Engineering
Marmara University
Istanbul/Turkey

***Abstract* — In this project we develop a classifier that separate the complaints written by bus passengers, to three different classes as Müspet,Uyarı and Menfi. These classes are determines if the driver should take a punishment or a warning or no punish required, respectively. But most of the complains are Müspet that means no punish required for driver and there are a lot of complaints that makes separation by humans very difficult. We use some machine learning algorithms to make this separation by computers and compare them to choose the best algorithm for this task**

Index Terms— Passenger Complaints, Classifying, Machine Learning Algorithms, Comparison

I. INTRODUCTION

IETT and Otobus A.S. are responsible institutions for public bus transportation in Istanbul. People use those buses may encounter some problems about bus, timing, driver, stop etc. So they can report the problems to Istanbul Municipality (IBB) via its “Beyaz Masa” channels. Our data is compiled from this reports from both civilians and personnel of IBB.

II. RELATED WORK

There are many works which use nlp algorithms and focus on classifying sentences written by people. One of them try to classifies the comments on internet as positive or negative to understand the quality of products that sold by various companies.

In another work, one of our group member has made a sentimental analysis on sentences written by twitter users and classifies these sentences as four different classes as fear, joy, anger and sadness. And he use a neural network algorithm to classify them and get an accuracy around %80.

III. APPROACH

Data format that we will work on is Comma Separated Values (CSV) with the columns labelled Durum (result) and Rapor Açıklaması (the text from reporter). There are 64314 records which have 3 different results (Durum) as Menfi, Müspet, Uyarı. The result labels are given to report records as result of the investigation and council decision. Müspet record means no punishment required and there is no guilty, Uyarı record means there is guilty but no punishment given for this time, Menfi means punishment applied to the driver.

In the CSV file we have 48702 Müspet, 14576 Uyarı and 3189 Menfi records. The council want to automate the elimination of Müspet records with a ML algorithm that trained by these records. The auto Müspet labelled records will be filtered before posting to the council.

Since the Rapor Açıklaması came from humans we need to employ some NLP algorithms. When we look at records we see some records have date data and some other not.

Some records we saw is about missing bus in a stop, bad behaviour from drivers, smoking driver, full-day working driver, driver passing stop, bus is not convenient for transportation etc.

IV. EXPERIMENT SETUP

Firstly we determine the number of complaints in each class and compute the frequency of each vocabulary in our data.

Then we turn our labelled data into unlabeled form to apply k-mean algorithm in order to saw if there is a natural relation between each complaints.

Lastly we try four different machine learning algorithm with three different dictionary sizes to train on data part that we separate from our dataset as %80 of it and then we calculate the accuracy of these algorithms by predict the classes of data part that we separate from dataset as %20 of it, by comparing them with real classes.

V. EXPERIMENTAL RESULTS

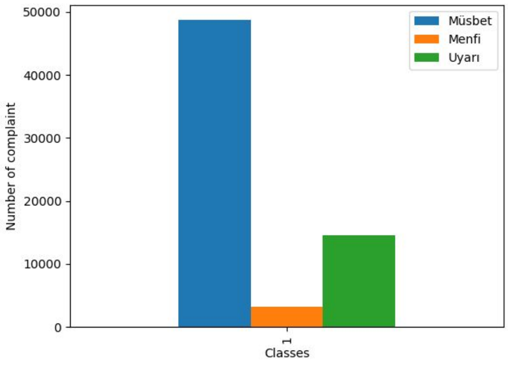

Figure 1: A bar chart that shows ratio of classes

In above graph we can see that most of the complaints are classified as Müsbet and very small amount of them is classified as Menfi. And some of them classified as Uyarı which can give penalty to the driver if he get another complaints.

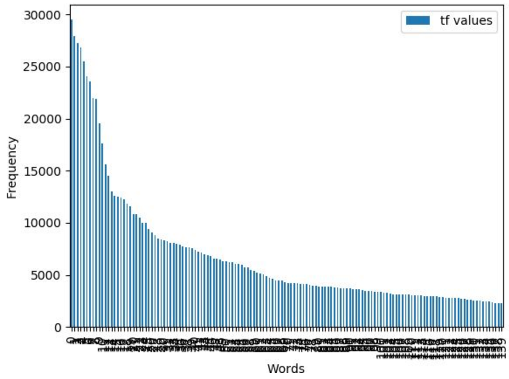

Figure 2: A bar chart that shows first 140 words with corresponding frequencies

In above graph we plot the graph of first 500 words with their frequencies and it is very compatible with Zipf's curve. It says some words repeated in most of the sentences in dataset and they don't help us to classify the sentences. This words are called as Stop words and we must eliminate them to get better results. Lastly we saw that only a small part of all words are in this Stop words class.

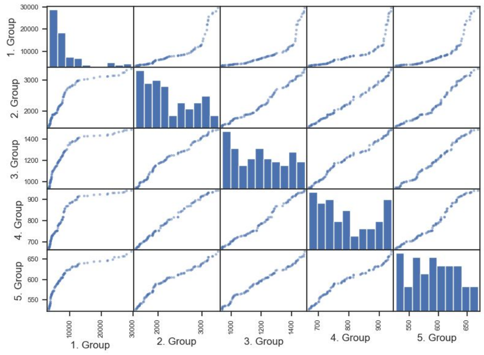

Figure 3:Scatter plot matrix of first 500 words group by 5

In above graph we divide the words and frequencies data into five groups so that each group contain a hundred words. Then we plot each pair in a five to five matrix which called as scatter plot matrix. In these lines we saw that the linearity decreases when one of the axis contains first group that contains Stop words whose frequencies are high valued and highly variable.

Then we use k-mean algorithm for clustering the sentences that already clustered by humans. To determine the ideal k value ,i.e. number of clusters, we test clustering algorithm with different k values then plot them with their distortion values that calculated with distortion function. When k value increases, the centroids are closer to the clusters centroids. At some point, the improvements will drop rapidly, forming the shape of the elbow and that point is the optimal value for k and this value is 3-4 which is very similar to real number of classes which are müspet,menfi and uyarı.

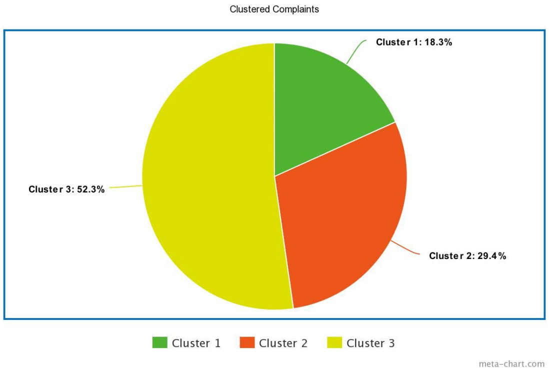

Figure 4: Pie chart of complaints in 3 cluster with percentages [1]

# of Clusters	Standard Deviation	Distortion
k=2	1197.53	0.1553
k=3	585.22	0.0970
k=4	210.93	0.0750
k=5	389.47	0.0566
k=6	462.54	0.0437

Table 1: Clustering experiments with different k values

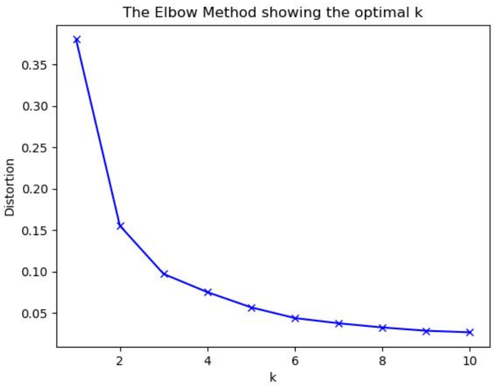

Figure 5: Different k values with their distortions

1. Learning algorithms and their results

1.1. Naive Bayes Algorithm

4138 sentences with 3 different classes

Max features 1000 (Dictionary size 1000)

Training set=%80 Test set=%20

Accuracy = 0.18478

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.845588	0.965517	0.0267857
F1 Score:	0.185022	0.066335	0.263923

Max features 2000 (Dictionary size 2000)

Training set=%80 Test set=%20

Accuracy = 0.18115

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.845588	0.965517	0.0267857
F1 Score:	0.185022	0.066335	0.263923

Max features 4000 (Dictionary size 4000)

Training set=%80 Test set=%20

Accuracy = 0.18719

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.0147059	0.965517	0.991071
F1 Score:	0.285106	0.06633	0.0176991

1.2. K Nearest Neighbor Algorithm

Max features 1000 (Dictionary size 1000)

Training set=%80 Test set=%20

Accuracy = 0.702898

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.926471	0.0137931	1
F1 Score:	0.12987	0.823022	0

Max features 2000 (Dictionary size 2000)

Training set=%80 Test set=%20

Accuracy = 0.700483

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.926471	0.0172414	1
F1 Score:	0.128205	0.821326	0

Max features 4000 (Dictionary size 4000)

Training set=%80 Test set=%20

Accuracy = 0.704106

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.926471	0.012069	1
F1 Score:	0.130719	0.823868	0

1.3. Decision Tree Algorithm

Max features 1000 (Dictionary size 1000)

Training set=%80 Test set=%20

Accuracy = 0.69927536

Class Statistics:	Menfi	Müspet	Uyarı

Miss Rate:	0.985294	0.005172	1
F1 Score:	0.0283688	.0.822523	0

Max features 2000 (Dictionary size 2000)

Training set=%80 Test set=%20

Accuracy = 0.69927536

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.985294	0.005172	1
F1 Score:	0.0283688 .	0.822523	0

Max features 4000 (Dictionary size 4000)

Training set=%80 Test set=%20

Accuracy = 0.69927536

Class Statistics:	Menfi	Müspet	Uyarı
Miss Rate:	0.985294	0.005172	1
F1 Score:	0.028368	0.822523	0

1.4. Neural Network Algorithm

Data Length: 2624 (complaints)

Train Part: %80 Test Part: %20

Accuracy: %79.8

→ We can't calculate f1 score

VI. CONCLUSİON

Our first algorithm was naive bayes algorithm and this algorithm gives the worst accuracy like 0.18, then as we increase the dictionary size the accuracy is increase very little but it was still the worst accuracy and most of its accuracy came from the 'Uyarı' class which is not least but very few in dataset.

Second algorithm was K-nearest-neighbor algorithm and its accuracy was over %70 which is far better that naive bayes algorithm. The increase in dictionary size also increase the accuracy very little again. And most of its accuracy came from the 'Müspet' class which is very much in dataset.

Third algorithm was Decision tree algorithm. Its accuracy is also better than naive bayes algorithm but very little worse than k-nearest neighbor algorithm and it was under %70. As we increase the dictionary size, the accuracy is doesn't change at all. Most of its accuracy came from 'Müspet' class and this algorithm may be the most overfitted algorithm because its miss rate in Müspet has the smallest value among all algorithms.

Our last algorithm was a neural network algorithm and it has the best accuracy which was 79.8. But we can't calculate the f1 score because the code was a little homemade code which didn't use a machine learning algorithm from library so functions that calculates f1 score are didn't work properly.

VII. REFERENCES

[1] https://www.meta-chart.com/pie, to produce py chart