# LDA+SVC Results

Hasan PARA, Elgun JABRAYILZADE, Algın Poyraz ARSLAN

November 2019

## 1 Results

In this experiment we try to understand the importance of the length of the document to the lda. That's why we take equal numbers of randomized documents from both city and disease. Then we split them to groups accordingly to their length. After that we fit lda model and linear SVM with each of those groups. Test those model the same test data. Given 19546 city document and 19546 disease document total document 39092. Those 39092 document split 17 groups. Each groups have 2175+-175 document. First graph show us the accuracy score of each group. Second and third graphs show us the city and disease document numbers of the each groups respectively.

Like the graph shows increase of the city document with decrease of disease document in groups increase accuracy and decrease of city with increase of disease decrease the accuracy score. Below graphs show us the accuracy score of each groups one document. First graph shows city accuracy and second graph shows disease accuracy scores.

In these graphs we see that almost all of groups accuracy score of the disease documents less than city document.

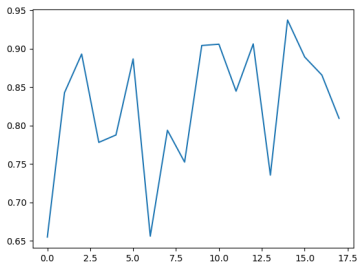


Figure 1: Data Groups Accuracy Scores

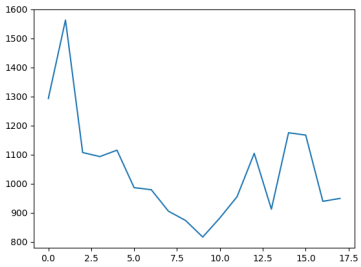


Figure 2: Number Of City Document In Each Groups

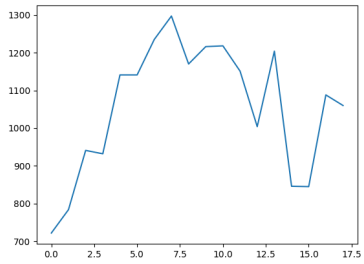


Figure 3: Number Of Disease Document In Each Groups

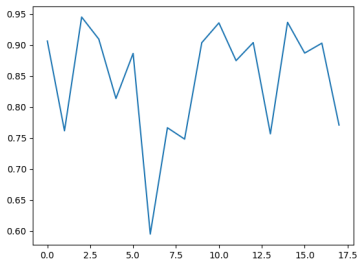


Figure 4: City Accuracy Scores According to Groups

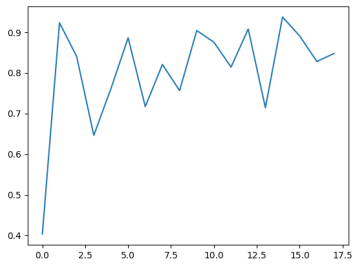


Figure 5: Disease Accuracy Scores According to Groups

If frequencies of the subtopics inspected in detail, in general, it can be seen in both data set, as the frequency of a topic is increases miss-classification rate decreases in the corresponding figure.

In consideration of figures 8 and 9, this generalization fits the figures roughly but some categories do not fit. For instances of 16th and 19th category in this data set, even though a considerable increase occurs in frequency corresponding miss-classification rates do not affect as dramatic as the frequency. The reason for such instances has not inspected in detail but the scarcity of unique words and common similar words are suspected.

The same case holds for the city data set as well. An increase in frequency generally decreases the miss-classification rate. On contrary to disease data set exceptions, even though no considerable amount of increase occurs a dramatic decrease can be examined for some instances of city categories such as 22nd and 24th. Opposite of the disease can be considered for the reason for that which is the plentifulness of unique words and the scarcity of similar words.

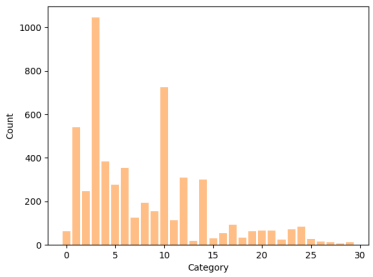


Figure 6: City topic frequency

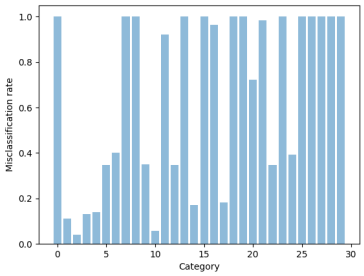


Figure 7: City topic miss-classification

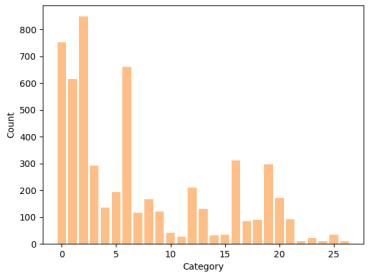


Figure 8: Disease topic frequency

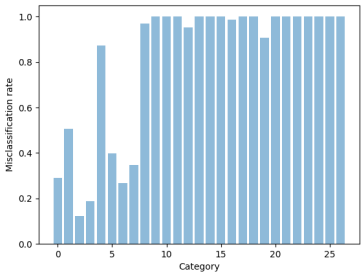


Figure 9: Disease topic miss-classification