Report for mandatory Problem - NLP

Matthias Thurnbauer
Deggendorf Institute of Technology
Deggendorf, Germany
matthias.thurnbauer@stud.th-deg.de

Ihab Ahmed
Deggendorf Institute of Technology
Deggendorf, Germany
ihab.ahmed@stud.th-deg.de

Karim Sayed
Deggendorf Institute of Technology
Deggendorf, Germany
karim.sayed@stud.th-deg.de

July 1, 2022

1 Introduction

1.1 Goals of our project

1.2 Importance of problem and why it has to be solved

1.3 Related Work

2 Methodology

2.1 Mathematical foundations of our model

2.2 Conceptual foundations of our model

3 Results

This section covers the results our model achieved and a description of our used dataset for training and testing the model.

3.1 Dataset description

We use a dataset consisting of 5572 ham and spam SMS messages which are not chronologically sorted [**dataset**]. We first randomize the whole dataset and then split it to 80% train and 20% test data, respectively, which results in 4458 entries for training data and 1114 for test data. After that we validate the splitting process by taking a look at the label distributions. We've found that we have around 86% ham and 14% spam messages in the training data whereas the test data has around 87% ham and 23% spam messages. We can conclude that this is a good data distribution for both datasets as there are a lot more ham than spam messages in the real world.

3.2 Metrics

3.3 Strength and limitations

3.4 Other interesting findings

4 Conclusion

4.1 Summary

4.2 Further improvements