Capstone Proposal
Sentiment Analysis for German Twitter

Ariel Brandes

August 12th, 2020

Domain Background

Nowadays the opinion on companies in social media is crucial for their success. Shitstorms arise frequently can have a huge impact. This makes it necessary for companies to monitor this sentiment on social media platforms to be able to react as fast as possible. To be able to extract sentiments from social media like twitter, well-trained machine learning models are needed. In general, sentiment analysis is classic problem in the area of natural language processing (NLP). Doing sentiment analysis on platforms like twitter has, however, its very own challenges. People use emoticons, hashtags and hyperlinks in their tweets, which makes it more difficult to decide which elements of the text are actually useful for the analysis and which ones are noise. NLP is very specific for the language it analyzes. German, my mother tongue, is a more complex language than English is and it is not as widely used as English is. This is the reason that less scientists concentrate their studies on the German language and less data and fewer benchmarks are available. It also results in fewer pre-trained models, which makes it necessary to train self-made models more often.

Problem Statement

Sentiment analysis is a classification problem. Each text (or tweet in this context) will be given a sentiment. For this aim, a tweet needs to be categorized into one of the three categories: positive, neutral, negative. Since the data for this project is labeled, the result is measurable and quantifiable. This problem can be tackled by neural networks or classical machine learning models. The approach of this project will be the latter which makes is necessary to have a good routine for data preparation and preprocessing, since tweets usually contain a lot of noise and parts that are not text like emoticons. In a nutshell, the models takes the text of a single tweet and proposes a sentiments (positive, neutral, negative) for this tweet.

Datasets and Inputs

The data set which will be used is a corpus of German tweets called SB10k. The tweets have been selected and categorized by hand in one of three sentiment groups (positive,

neutral, negative). The corpus was introduced by a group of scientists [1] and originally featured around 10,000 tweets. The corpus is publically available and hence a good source for projects and resources. Since the authors were only allowed to provide a list of tweets id (they also offer a script to download the tweets), not all of the originally used tweets are still available. Nevertheless more than 6,500 tweets are still there which is enough for most machine learning models.

Solution Statement

The problem is clearly a supervised learning problem, since the dataset is labeled with a sentiment (positive, neutral, negative). The goal is to predict a sentiment for a text originating from a tweet. This implies the use of a classification model. Since the model gives a prediction for each (labeled) tweet, the solution will be quantifiable and measurable. Accuracy of the model can be calculated as well as F1 scores.

Benchmark Model

The authors of the paper, which published the corpus of German tweets [1], also did sentiment analysis and provided benchmarks of a convolutional neural network (CNN) and of a feature-based support vector machine (SVM). For the CNN, they also created German word embeddings trained on 300M tweets. These word embeddings were then optimized for sentiment analysis using distant-supervised learning. For the corpus used in this project, they were able to achieve an F1 score of 56.98 for the SVM and an F1 score of 65.09 for the CNN. They also provide specific F1 scores for each sentiment (positive, neutral, negative).

Evaluation Metrics

As mentioned above the actual accuracy (correct predictions/all predictions) will be used as well as the F1 score as my evaluation metrics. The F1 score is a widely used measure for accuracy taking the precision and the recall into consideration. This metric also allows me to compare my results to the results achieved in the paper.

Project Design

The first step, like in every machine learning project, is to explore the data. Downloading the tweets from the original corpus will show how many tweets are still available. Subsequently the data will be cleaned in order to get rid of the tweets that are not available anymore. In order to gain further insight, the data will be visualized. Plotting the count of each sentiment label will provide a good overview of the distribution of the labels positive, neutral or negative, respectively. Since this is a NLP problem and tweets usually contain a lot of noise, the next step is preprocessing the data. For this aim different techniques will be used including removing hyperlinks, removing hashtags, removing emoticons, tokenization, correcting typos, stemming, lemmatization, removing stop words. Not all of the named techniques have proven to be helpful for all machine learning based sentiment analysis, so

it will be evaluated which combination of these seems most promising. The next step is dividing the data into training and testing set. It is planned to train different models using standard parameters in order to find the one best performing. The models planned to be used are random forest and SVM. The models will be evaluated based on their results and on computation time. As mentioned previously, the most important metric will be the F1 score. After this, the best models will be trained again with differently preprocessed data to investigate whether altering the preprocessing step can improve results. The final step is to modify hyper parameters so further improve the result.

References

[1] Cieliebak et. al. (2017) *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*, Association for Computational Linguistics, Valencia, Spain.