Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the Lecture Text Analytics

Time-dependent Sentiment Analysis for Covid-19 including a Correlation Analysis to Relevant Political Events

Team Member: Sina Denzel, 4018461, Computational Linguistics B.A
sinadenzel@gmail.com
Team Member: Ahmad Fadlallah, 3442106, Applied Computer Science B.A.
abohmaid@windowslive.com
Team Member: Ute Gradmann, 4050818, Computational Linguistics B.A.
utegradmann@gmx.de
Team Member: Severin Laicher, 3665790, Applied Computer Science M.Sc.
severin.laicher@web.de

1 General

The code of our project will be available at: https://github.com/Ahmad-fadl/ita.

As for the data, we are planning to use the Coronavirus geo-tagged tweets dataset.[1] The dataset contains about 270 thousands of English corona related geo-tagged tweets in the form of their tweet-ID. The dataset consists of one csv file per day, starting on March 20th until today. Each entry consists of exactly one tweet-ID and a sentiment score (which we ignore). The dataset is described in [1]. Since our project aims to analyze a sample of corona-related tweets from different time periods, this dataset is suitable.

2 Motivation

Analyzing public opinions on certain political topics using textual data is one key area of text analytics. By this mean, the people's needs and interests can be investigated and the results can possibly lead to governmental policies. A current example is the process of Covid-19 infections. The last year has shown very divergent opinions about how to deal with the Covid-19 pandemic and which measures to take to protect society and economy.
The motivation of our project is to put the political decisions and measures as well as the infection process in relation to twitter data. First, we might discover a potential correlation between the number of tweets and the number of infections or political decisions. Second, a sentiment analysis can reveal the public's opinion on certain measures. Do people find the measures taken appropriate? Which ones lead to positive or negative expressions? To which extent are people willing to follow political decisions on personal restrictions? An analysis of this kind is a helpful tool for the government as they can directly assess the public opinion. On the one hand, assessing the public's opinion could prevent intensification and extremism by changes in the communication strategy. On the other hand, politicians might want to gain votes and therefore find out what the people want.
As different countries have pursued different strategies regarding the Covid-19 pandemic, in the beginning we want to focus on one specific country. If we still have enough capacities, we additionally consider a cross-national comparison of two or more countries, in order to detect differences or similarities in the population's opinion.

[1] `https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset`

3 Research Topic Summary

Sentiment in twitter data has been analyzed since the late 2000s.[3] Twitter data analysis has also been conducted for tweets on Covid-19.[1]
Sentiment analysis is a subfield of text analytics and aims to categorize the sentiment of a given text or phrase. For this purpose, different aspects of the text can be analyzed on a lexical and structural level. The most obvious approach is to use a lexicon with words categorized in different emotions, be it coarse grained positive or negative, or more fine grained related to specific emotions like joy, hate, excitement etc. Still, some further aspects need to be taken into consideration when classifying the sentiment. One aspect is negation detection and structural elements like the presence of personal pronouns or the amount of punctuation are markers for the intensity of subjectivity. Our approach will be based on the emotion lexicon[2] developed by Mohammad and Turney 2013. [2] The words of a tweet will be compared to the entries in the NRC Word-Emotion Association Lexicon and the tweet categorized accordingly. In addition, we will detect negation and evaluate the level of subjectivity indicated by the amount of personal pronouns and punctuation marks per tweet.

4 Project Description

4.1 Tasks

As a method-driven approach we want to analyse sentiment (nearly) from scratch and implement some methods of the lecture, to see how well a coarse sentiment analysis will work. The task in this project is to establish a text analytics pipeline that automatically extracts and analyses Covid-19 related tweets. We want to demonstrate how well social media represents the nationwide opinion regarding a controversial topic. In particular we want to investigate the following tasks:

1. **Theoretical Work**
   We would like to find correlations between political measures regarding Covid-19 and the twitter activity. Therefore, we choose a specific country (USA) to find out about the most important measures there, in order to create a timeline.

2. **Sentiment Analysis**
   Create sentiment analysis for Covid-19 related tweets from the three

[2]https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

countries at one specific period of time by implementing our text analytics pipeline, as described below.

3. **Time-dependent Sentiment Analysis**
   So far we have only performed a sentiment analysis for a fixed point of time. The next step is to repeat this for multiple points of time, in order to create a time-dependent sentiment analysis. We then can plot the developments in sentiment and infections rates over time.

4. **Evaluation of Sentiment Analysis**
   We plan to create a test set in order to evaluate the performance of our sentiment analysis. The test set and the evaluation is described below.

5. **Correlation Analysis**
   As mentioned we are searching for correlations between political measures and the twitter activity regarding Covid-19. Therefore we will analyse the data to see if there are connections with the previously created timeline.

6. **Cross-national Comparison**
   As different countries have pursued different strategies regarding the Covid-19 pandemic, we mainly want to focus on one specific country (USA). If we have enough capacities, we additionally consider a cross-national comparison of two or more countries, in order to detect differences or similarities in the populations opinion. That would be USA, New Zealand and Great Britain, since those countries pursued different Corona strategies and are all English-speaking.

4.2 Sentiment Analysis Pipeline

1. **Target data selection**
   As mentioned below we use an already complete dataset of tweet-IDs, therefore the target data selection step of the KDD process is already applied.

2. **Text preprocessing**
   After extracting the tweets for the given tweet-IDs, the text will be processed by applying the following steps.

   (a) Tokenization

   (b) Normalization

   (c) Stopword removal

3. **Text transformation**:
   In order to assign each tweet either the sentiment label *positive*, *negative* or *neutral*, the text will be transformed into a vector representation. To do so, we first assign sentiment to words in the documents (Tweets) using an emotion lexicon, taking special punctuation (e.g. "!!!!") or special writing of words (e.g. complete word in capital letters) into account.

4.3 Data

As mentioned above, we plan to use the Coronavirus geo-tagged tweets dataset.[3] Considering the knowledge discovery in databases process we do not start from scratch, but save time by using already pre-selected data. The mentioned dataset perfectly fits our project goals, since it contains about 270 thousands of English corona related geo-tagged tweets in the form of their tweet-ID. The dataset consists of one csv file per day, starting on March 20th until today. Each entry consists of exactly one tweet-ID and a sentiment score (which we ignore). The fact that the dataset only contains geo-tagged tweets is very import for our project, since we want to assign each tweet to a specific country. In general this is not given for all tweets, since the user has to explicitly agree that his tweets get geo-tagged. We will extract the tweets associated with the corresponding tweet ID using the Twitter API. [4]

4.4 Evaluation

The first part of our project is searching for a correlation between number of tweets and number of infections. This will be measured by calculating the Pearson correlation. Additionally, we investigate whether the main political measures regarding Covid-19 have any influence on the distribution of the sentiments in the corresponding country.

One challenge to trace multiple sentiments at different points in time. Therefore it is important to verify the correctness of our sentiment analysis model. Another challenge is that we do not know the ground truth sentiment scores for the tweets. We therefore decided to manually annotate a sample of tweets (200-1000 tweets) of appropriate size. The subset of tweets is selected by applying systematic sampling, since we do not know any subgroups for applying stratified random sampling. We will split the sampled tweets by four and each

[3] `https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset`
[4] `https://python-twitter.readthedocs.io/en/latest/`

member of our team will label a subset of the tweets manually. This results in our ground truth test set on which we will evaluate the performance of our sentiment analysis model by calculating precision, recall and accuracy.

4.5 Project Timeline

December

- Data acquisition
- Research of political events
- Developing methods for sentiment analysis

January

- Text preprocessing
- Data collection and analysis

February

- Data analysis and interpretation
- Project presentation
- Project report

March

- Finalizing report

References

[1] Rabindra Lamsal. Design and analysis of a large-scale covid-19 tweets dataset. *Applied Intelligence*, pages 1–15, 2020.

[2] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[3] Mike Y. Chen Sanjiv R. Das. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9), 2007.