Technology
Arts Sciences
TH Köln

University of Applied Sciences Cologne
Deep Learning Algorithms

Research Project - Proposal

Objective

The objective of this research project is to design and develop a neural network that is able to create and publish short text messages, further named tweets, on the social networking service Twitter. The content of these tweets will be based on the tweets of a public personality. Since there is a steady increase of machines and algorithms in our everyday life, researches are looking for ways to facilitate human machine interaction. The most natural way for humans to communicate with other beings is spoken and written language. Therefore, on the one side, it is essential to develop algorithms, which are able to extract information from human language. On the other side it is not less important that these algorithms can form sentences, which are understandable for human beings.

Related Work

Sentiment analysis in the domain of micro-blogging is a relatively new research topic so there is still room for further research in this area. A decent amount of related prior work has been done on sentiment analysis of user reviews , documents, web blogs/articles and general phrase level sentiment analysis . These differ from twitter mainly because of the limit of 140 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines. However the manual labeling required for the supervised approach is very expensive. Some work has been done on unsupervisedand semi-supervised approaches. Various researchers testing new features and classification techniques often just compare their results to base-line performance.

Technical Outline

As a first step the tweets need to be gathered from a chosen public person. For this purpose Twitter offers an API, which is wrapped in several Python packages. The extracted data will be stored in an online database. The second step is to preprocess the data in a meaningful format, so the neural network can train optimally. The disturbances in Tweets can be really high, so the procedure must be well thought. The tweets can include misspelling, hashtags or links, which will make it difficult to learn on the dataset. In the third step the data will be given to a neural network. Because of the condition to have a variable length input and to share learned features across different positions of the text, a standard neural network cannot be selected. A possible choice will be a neural network of the RNN family. The base principle of RNNs is the recurrent hidden state, whose activation is dependent on the hidden state of the previous time. With this the previous decisions can influence the next decisions, which is needed to learn features across a sentence. A disadvantage of the RNN is the problem of the vanishing gradient. This affects the longterm dependencies, which do not contribute sufficiently to the decision making process of an RNN. To solve this problem, there are GRU (Gated Recurrent Units) or LSTM(Long Short Term Memory), which adds the possibility to contribute longterm dependencies to the current deci- sion, because of the introduction of a cell state, which can carry longterm information.