Prediction of Comments in Online Media

Degree programme : BSc in Computer Science | Specialisation : Data Engineering
Thesis advisor : Dr. Souhir Ben Souissi
Expert : Pierre-Yves Voirol

In this bachelor thesis, a model is trained with the help of machine learning using article texts from a news portal and the corresponding comments. The model predicts comments that might be written on an article by the portal‘s community. As a result, the community‘s reaction to a news article can be predicted.

Introduction

Comment sections have their advantages in many areas. Comments can help an author see errors in his posts and get feedback from his readers. News portals allow readers to express their opinion on a topic and discuss it with other readers. A user base can be created, which regularly returns to the portal and interacts with the site. For an online portal, this is important, as most are financed by advertising or subscriptions.

Goal

Through the comment sections, the mood of a community to a specific topic can be seen. In this thesis, machine learning was used to train a model to predict reactions from a specific community by generating comments on a text.

Contribution

It was decided to use the Watson community for this thesis during the preliminary work since the comments were the most useful and least offensive. During prior work, German article texts and the corresponding comments were collected from the news portal Watson.ch. A pre-trained GPT-2 Transformer model was fine-tuned with the collected articles and comments to learn to predict the reaction from a specific community by generating a comment on a given text. A method was successfully tested in which an open-end text generation can distinguish between article and comment by separating the two texts with tags. The comment is written in a different vocabulary and style than the article and responds to it.

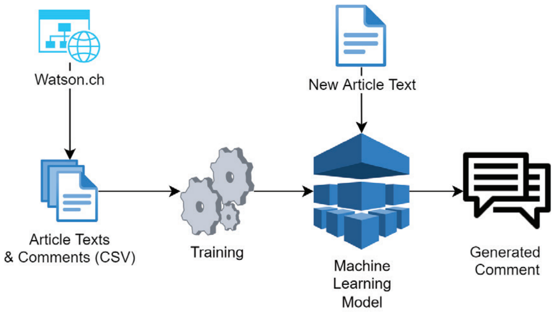
Results

It quickly became apparent that it is essential for text generation that stop words are present. The model otherwise learns to write in very incomprehensible grammar. Additionally, when training with all data, it became apparent that the resources were eventually overloaded with this amount of data, which caused the training process to crash. Therefore, the process was rebuilt to train the model multiple times with a smaller batch of data and save the current learn state as a checkpoint. This checkpoint was loaded again for the training with the next batch of data.
Another problem was the evaluation of the model, as it proved difficult to evaluate the grammar and vocabulary. The model has learned to work with tags, leading many evaluation metrics to rate the results as poor, as the generated text seems incomprehensible. Therefore, the sentiment of the comments was evaluated. A process was created to estimate the sentiment of the generated comments and the comments on the article. These are compared afterward to evaluate the model.

Conclusion

These results show that it is possible to train a model which can respond to a text in the community‘s view. The model has learned to write in the community‘s style and vocabulary. In addition, using tags to distinguish comment text from article text turned out to be promising: The model can differentiate between article and comment and changes the style and vocabulary each time the text is generated.

Nico Lieberherr


Watson.ch
New Article Text
Article Texts & Comments (CSV)
Training
Machine Learning Model
Generated Comment