Online content moderation service with machine learning engineering

Degree programme: BSc in Computer Science | Specialisation: Data Engineering
Thesis advisor: Prof. Dr. Erik Graf
Expert: Reto Trinkler

There are millions of comments posted to online message boards every day and platform providers spend a huge amount of resources to moderate them according to the law. This work tries to ease the situation by creating a web service that can make moderation decisions on its own by applying machine learning engineering. The focus is to produce a robust and user-oriented service that considers all aspects of a practical real world product.

Context and Goals

More and more the public square is not the place where people go to communicate ideas and opinions. Today, online communication platforms like Twitter, Reddit and other user content forums are used to make up a large part of people's daily conversations. These places face the same problem as their real life counterparts, namely defining what should be allowed to be expressed. However, the technology of the internet brings many additional challenges like the enormous amount of content being produced every day. The idea of this thesis is to tackle this problem by applying machine learning and data engineering techniques.
The goal of this thesis is to create a web application that can evaluate comments and give an opinion on whether the comment should be moderated or not. This application should not just focus on one aspect like for example creating a perfect machine learning model for evaluation, but it should instead go through the entire machine learning engineering cycle from data collection to maintenance.

Methods

The thesis is split in two parts, a practical and a theoretical one.
For the practical part, a web application was made by implementing every step of the machine learning engineering cycle.
In the theoretical part, I documented problems and questions about design decisions that came up during the practical part and explored them in detail, giving solutions and showing aspects and thought processes.
The process was structured in 6 general topics:

- API with documentation and user interface
- Machine learning model selection and data pipeline
- Feature engineering
- Interpretability and user feedback
- Testing and comparison to market solutions
- Deployment and monitoring

Results

The goals of the thesis have been met. The resulting web application may look simple on the surface, but the features under the hood are well thought out and could be used as a basis to scale up to a real product. The application provides an API for platform providers as well as a simple graphical frontend to demonstrate the service. The moderation decisions themselves have a confidence value and a specific reason for the decision provided to help a user trust the decision. There is also a feedback labeling tool that can be used to improve the machine learning model over time and a documentation of the API.
The moderation decisions are not extremely accurate mostly due to a lack of training data, but the accuracy is high enough to be usable and by putting a focus on interpretability and user feedback, decision mistakes are easier to spot, and the model should get more and more data in the future, increasing the quality of the service over time.

Conclusion

Implementing a service like this in a real world product is a huge task and each of the covered topics potentially merits its own thesis. Therefore the project was realized by focusing on a small scale, and the depth of exploration for specific aspects was sometimes limited. A key insight of this work was, that solving online content moderation requires more than just letting a machine learning model do the decisions, but it can be very beneficial, if done with great care. The most interesting aspects of the whole project was definitely the interpretability and user feedback. These two aspects really showed the importance to connect every single aspect of the service back to the user experience. Rather than spending weeks perfecting a machine learning model to get 1-2% more in some testing metric, the focus was placed on all major aspects and how they relate to the big picture of creating a practical solution for the end user.

Simon Caspar Sterchi