Toxic Comment Identification and Classification using Deep Learning

ML Internship Team 7.7

Abstract

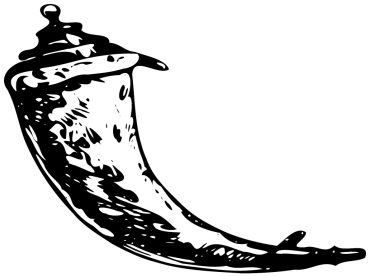
Every day a tremendous amount of data w.r.t to video, audio, images, and text from various social media platforms are generated. Out of which text data is majorly generated in large volumes. This text data contains toxic comments like abuse towards a person community or some threat to some person or a community that can be severe if not identified early and eventually will take an ugly turn. So to address such issues, the severity of toxic statements needs to be identified so that conscious action against such toxic people could be taken. Since the volume of text data is huge we cannot monitor it manually and the ambiguous nature of threat patterns could even be unidentifiable. Social media has a direct or indirect effect on our mental health and toxic comments can worsen mental health. Toxic comments can cause personal attacks, online harassment, and cyberbully, serious risk to mental health, and emotional health. Sometimes toxic sentences are used to cause communal disharmony which has worse effects on society. In this era of ever-growing online discussions, learning sources, and social media have increased toxicity in the form of online comments and discussion forums. Toxic comments spread hate and it’s ever-growing due to contrasting views of different groups of people covering up under some anonymous username can affect mankind mentally, emotionally, and make them stressed out a lot that they can’t even focus in their life. We are developing this project to help maintain social authenticity. We are exploring various neural-based and NLP methods to build an accurate model that can detect diverse types of negative online comments perceived by users or people.

Technology Stack

- Deep learning framework used Keras and Tensorflow
- Programming Language used is Python
- Basic Web application (If time permits)
  - HTML, CSS or maybe react based.
- API Development using Flask

TensorFlow

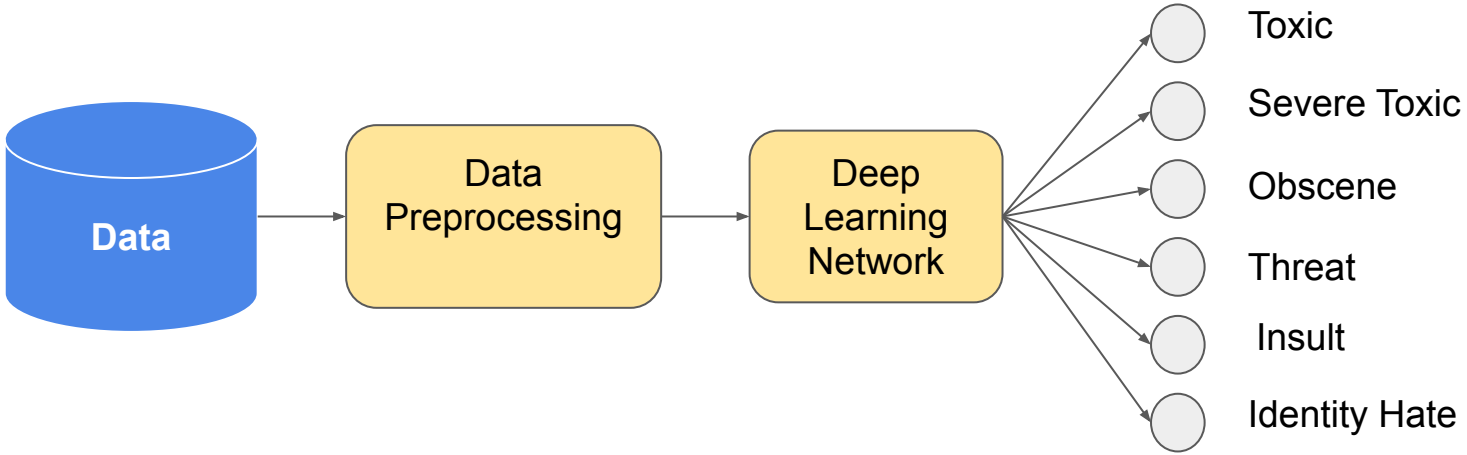
Keras



Flask

python™

Project Design and Implementation



Project Design and Implementation

- The first step is data gathering, we gathered data from kaggle source.
- After that we applied various data preprocessing techniques such as punctuation mark removal, redundant space removal, converting all words into lowercase, etc.
- After this we tokenized the data and used GLOVE embedding vectors for converting into continuous space vectors.
- Once this is done we are experimenting on various deep learning based architectures for efficient identification and classification of toxicity.
- Majorly we are using keras framework for our deep learning architecture building.
- We also aim to put it in a form of api using Flask framework.
- If time persists we will try to build a simple web application.

Conclusion

By building this neural-based system to identify toxic statements, we can track various social media comments which would worsen the communal harmony, a threat to a person, or maybe lead to various social issues. This can help create healthy communication on various social platforms which has many benefits.