MULTILINGUAL ABUSIVE LANGUAGE DETECTION USING MACHINE TRANSLATION

GitHub Repo link: https://github.com/asim-cyb/multilingual-abuse-detection/

Team Members:

Project Group 9

Soumya Kodumuri (11602632)

Sai Varun Gadde (11634194)

Karthik Amarnath Cholleti (11605154)

Asim Mahroos Mohammed (11558526)

MOTIVATION

It is critical to detect abusive language in order to create a secure and inclusive online environment. But because an increasing number of people are using more than one language to talk online, it can be difficult to find abusive language in different languages. Here, machine translation can come in handy.

By using machine translation to find abusive language in more least 1 language, we can make a more precise system that can find abusive language even if it is written in a different language. This can be very important in online communities, where verbal conversations in different languages.

Moreover, machine translation can help with the issue of low-resource languages, where there may not be enough training data to build an accurate system for detecting abusive language. We can improve the accuracy of detection in languages with few resources by using machine translation.

There are many other reasons for using multilingual abusive language detection by machine translation.

- We can build better communication by identifying abusive language across language barriers, online platforms can facilitate better communication between users who may not speak the same language.
- Create safe environment by detecting and removing abusive language in online platforms which creates a safer and welcoming environment for users which helps to reduce the incidence of online harassment, and encourage more constructive and respectful communication.

- Brands can improve their reputation by promoting a safe and inclusive online environment. It can improve their brand reputation and enhance their image as a responsible corporate citizen. This can help to attract more users and customers and can lead to increased revenue and growth.

Overall, using machine translation to find abusive language in multiple languages can help make the Internet a safer and more welcoming location for everyone, especially in multilingual communities and low-resource languages.

OBJECTIVE

With the growth of social media, people from different countries have started social networking. As people from different countries use social media, variation in language is vast and people communicate in their own language. Users from different age group started using the internet. Even the kids of age 10-15 are good at using the smartphone. So, hiding abusive language has become more important as it would effect the kids and most of the people would like the hate and trash talk away.

By building this project, we are planning on detecting abusive words from almost all the languages. The objective of multilingual abusive language detection using machine translation is to instantly detect and flag occurrence of abusive language across different languages.

By using machine translation, we are planning on converting the text coming from different languages into most commonly used language which is English. By translating, single model would be enough to detect the abusive language. Without using different abusive detection models for different languages, which requires a lot of datasets, we will only use single model with the help of machine translation which would make the detection system more efficient.

Tech-Stack:

For the machine translation, we are planning on using open source tools like google translate and Microsoft translate

For the text extraction, we are planning on using NLP libraries of python like NLTK and Spacy.

For the detection of abusive language, we are planning on using deep learning models like CONVOLUTIONAL NEURAL NETWORK (CNN). To train and develop these models, we will be using libraries of python like PyTorch and Keras.

For testing the accuracy, we will be using multiple evaluation methods like confusion matrix, ROC curve and precision score.

Workflow:

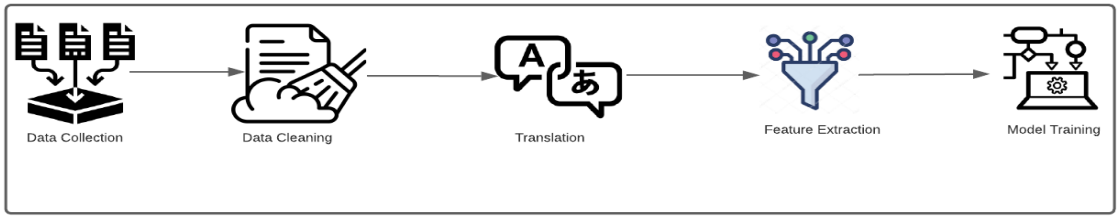

Fig: Workflow Diagram

The above diagram would describe our basic workflow and overview of the project. Other processes or libraries might be added in the future.

SIGNIFICANCE

Using machine translation to find and stop hate speech, cyberbullying, and other forms of harmful language in different languages and cultures is a powerful way to find and stop hate speech. Here are some of the most important things about this method:

1)Greater precision: By allowing text in more than one language be analysed, machine translation can assist enhance the precision of multiple languages abusive language detection. This can help look for patterns and ways of speaking that are specific to different cultures and languages. This makes it easier to find abusive language.

2)More Coverage: Detecting abusive language in multiple languages can help find inappropriate content in translations that aren't always well-represented in data sets and models. This can help make sure that efforts to find and moderate content don't just focus on a small number of languages but include a wide range of languages.

3)Quick response: Leveraging machine translation to rapidly translate abusive content from one language to another can decrease the amount of time it takes to respond to abusive language in real time. This is especially important on social media platforms and other online forums where negative content can spread quickly.

4)Abusive detection: Using machine translation to find abusive language in multiple languages and cultures can help raise awareness of cross-cultural differences and raise understanding between different groups.

5)Better Moderation: Multilingual abusive language detection can be used to improve content moderation by automatically finding and flagging abusive content in multiple languages. This can reduce the burden on human moderators and create online communities safer and nicer overall.

FEATURES

Multilingual abusive language detection using machine translation involves translating the text into a common language, then using machine learning models trained on abusive language in that language to identify abusive language.

This is how it goes:

- A common language translation is done first. This common language can be English or any other language with a sizable dataset of examples of abusive language that have been labeled, as well as models for high-quality machine translation.
- Abuse is detected in the translated text using machine learning models. These models were developed using a sizable dataset of examples of common language abuse that had been labeled.
- The text can be translated back into the original language for the user to read once the offensive language has been located.

This approach has some limitations. One major limitation is that machine translation can sometimes be inaccurate, especially for languages that have complex grammar or idiomatic expressions. This can result in inaccurate or incomplete detection of abusive language.

In addition, this approach may not be effective for languages that have limited labeled data for training the machine learning models. This can result in lower accuracy and performance of the detection system.

Overall, multilingual abusive language detection using machine translation can be a useful approach in certain situations, but it is important to be aware of its limitations and use it in conjunction with other approaches to ensure accurate and comprehensive detection of abusive language.

In the future, depending on how much time is left to complete the project, we are willing to add more features.

REFERENCES

- https://link.springer.com/article/10.1007/s00779-021-01609-14
- https://www.sciencedirect.com/science/article/pii/S1319157821001804
- https://aclanthology.org/R19-1132.pdf
- https://arxiv.org/pdf/2207.06710.pdf

CONTRIBUTIONS

Soumya Kodumuri - Motivation

Sai Varun Gadde - Significance

Karthik Amarnath Cholleti - Features

Asim Mahroos Mohammed - Objective