Classifying cyberbullying on the Twitter dataset

Raul Rangel Moraes Bezerra

I. Dataset

The dataset I used originates from the article SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection [1]. It contains over 47000 tweets, which were manually classified into the following cyberbullying categories: age, ethnicity, gender, not cyberbullying, other forms of cyberbullying and religion.

II. Classification pipeline

A. Pre-processing

Before creating the classification pipeline, a few pre-processing steps were taken:

a) Removing mentions: Since mentions don't bring any value to the classifier, I decided to remove any word that starts with the character '@'. In order to do so, I used Python's regular expressions library [2] to modify the tweets and remove any mentions.

b) Lemmatization: I used Lemmatization [3] to convert words to their dictionary form.

B. Creating the pipeline

I used Logistic Regression [4] for the classification pipeline. The coefficients represent the weight of each word, aligning with the bag-of-words strategy by preserving the relevance of each word's meaning.

III. Evaluation

```
age
-> schools, bullied, bullies, bully, school

ethnicity
-> coon, dumb, colored, nigga, nigger

gender
-> female, sexist, notsexist, rape, feminazi

not_cyberbullying
-> daesh, mosul, andre, beatdown, mkr

other_cyberbullying
-> harassment, code, bullied, idiot, blameonenotall

religion
-> muslims, mohammed, islam, muslim, christian
```

Fig. 1. Top five words for each class of cyberbullying obtained from our classifier

The classifier was run 10 times, with the data shuffled each run. The accuracy score of each pipeline was stored along with its corresponding pipeline. Then, I took the pipeline whose accuracy was closest to the average accuracy of all pipelines as to avoid overperforming or underperforming pipelines and obtain a more reliable representation of the model's performance.

Here's an analysis of the results shown in Figure 2.

a) Age: All top words are strongly linked to bullying in schools, which indicates that age-based cyberbullying usually occurs in educational environments.

b) Ethnicity: Most words are slurs aimed at African-American people, reflecting ethnicity-based cyberbullying.

c) Gender: Words like female and sexist indicate possible gender discrimination, while more extreme words like feminazi and rape indicate misogynistic behavior.

d) Not cyberbullying: Results don't point to anything.

e) Other cyberbullying: Harassment and idiot may suggest types of cyberbullying not included in the dataset.

f) Religion: All top words are related to religion, but are not related to cyberbullying specifically. This could indicate a bias in the dataset, since harmful are the majority.

IV. Dataset size

After using downsampling [6] to observe the train and test accuracy curves from my model, I observed that the test accuracy stagnated. This indicates that in order to see a significant increase in accuracy, it would take a big effort in data collection. Given that tweets are abundant, this is technically feasible. However, it will require a lot of manual labor to label the new tweets.

V. Topic analysis

I used Topic Modeling with NMF [5] to identify distinct topics in the dataset and test classification performance across different topics. The results showed varying accuracy rates.

```
Topic 0 - Accuracy: 0.93 - words: (high bullied school girl girls got middle bullies mean friends)
Topic 1 - Accuracy: 0.94 - words: (fuck dumb nigger ass obama bitch niggers shit fucking black)
Topic 2 - Accuracy: 0.96 - words: (rape gay jokes joke funny people make making http men)
Topic 3 - Accuracy: 0.78 - words: (people don muslims muslim idiot black white know idiots christian)
Topic 4 - Accuracy: 0.87 - words: (bully school bullies middle high kids people used bullying kid)
Topic 5 - Accuracy: 0.65 - words: (rt http mkr sexist women obama https ass kat men)
Topic 6 - Accuracy: 0.77 - words: (like look looks shit feel bitch ur act make don)
Topic 7 - Accuracy: 0.73 - words: (just mkr female bitch right got really mean want sexist)
```

Fig. 2. Topics identified along with their accuracy and most relevant words

As seen on Figure 2, classification effectiveness varies significantly across topics. In order to achieve these results, I used a two-layer classification system. First, documents are classified according to their topic. Then, they are redirected to a specialized classifier fine-tuned for that specific topic. This approach allows the model to account for topic-specific nuances and improve performance where it is otherwise less effective.

REFERENCES

[1] J. Wang, K. Fu and C. -T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 1699-1708.

[2] RE - regular expression operations (no date) Python documentation. Available at: https://docs.python.org/3/library/re.html (Accessed: 03 October 2024).

[3] Gillis, A.S. (2023) What is lemmatization?: Definition from TechTarget, Enterprise AI. Available at: https://www.techtarget.com/searchenterpriseai/definition/lemmatization (Accessed: 03 October 2024).

[4] What is logistic regression? - logistic regression model explained - AWS. Available at: https://aws.amazon.com/what-is/logistic-regression/ (Accessed: 03 October 2024).

[5] What is logistic regression? - logistic regression model explained - AWS. Available at: https://aws.amazon.com/what-is/logistic-regression/ (Accessed: 04 October 2024).

[6] What is downsampling? (2024) IBM. Available at: https://www.ibm.com/topics/downsampling (Accessed: 03 October 2024).