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 [referencia], which can be found on Kaggle [referencia]. 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 [referencia] to modify the tweets and remove any mentions.

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

B. Creating the pipeline

I used Logistic Regression [regressao] 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

The classifier was run 10 times, with the data shuffled each run. The accurary 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.

I will now give a brief analysis the results of each class of cyberbullying as shown in Figure 1.

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 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

Topic

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.