Hate Speech and Offensive Language

Tuxun Lu, Ting Li, Keyi Ding
tlu32, tli79, kding5

Abstract

We propose the study training a machine learning model to identify hateful and offensive languages on social media. We will use a neural network model including RNN and LSTM.

1 Project choice

Choose either a **methods** or **applications** or **suggested** project, and a subarea from the below table.

☒ **Applications**

☐ Genomics data ☐ Healthcare data ☒ Text data ☐ Image data ☐ Finance data

☐ **Methods**

☐ Fairness in ML ☐ Interpretable ML ☐ Graphical Models ☐ Robust ML ☐ Privacy in ML

☐ **Suggested projects**

☐ Music AI ☐ Genomics & Bioinformatics ☐ NMT

☐ Adversarial ☐ Domain adaptation ☐ ML fairness

2 Introduction

Hateful and offensive languages are common on social media. While some social media has already published policies to prevent such undesirable behavior, it is still necessary to use automated method to accurately, effectively, and efficiently identify hate-speech and offensive languages. We hope to accomplish this goal using a machine learning method.

The input to our algorithm is an English sentence drawn from major social media such as Twitter. We clean the data to extract useful text information, and use a RNN neural network to determine whether the input sentences contain hate speech or offensive languages.

3 Dataset and Features

We will use dataset from Hate Speech and Offensive Language (Davidson et al., 2017), that contains 24784 sentences from Twitter and labels on whether people think the sentences are hateful or offensive. Since the input data is currently noisy, we first preprocess the data to remove non-text information (such as emojis and usernames), and extract only the useful texts from the sentences. We then use the word2vec algorithm to convert the sentences into numerical vectors (Mikolov et al., 2013). Finally, we use a stratified train-test split with a ratio of 7:3 to separate the word vectors into a training set and a testing set.

Here are some examples from the data set:

tweet	hate speech	offensive language	class
RT @TorahBlaze: @1SonofYahweh they should be ashamed of themselves. I'll be single for life before I fuck w a nasty, faggot ass white man.	3	0	hate speech
Any person that puts anybody else's business out there for whatever reason is a bitch.	0	3	offensive language
"@DomWorldPeace: Baseball season for the win. Yankees" This is where the love started	0	0	neither

Each tweet is reviewed by 3 users, and the hate speech and offensive language columns are numbers of people (out of 3) who judge the tweet as hateful/offensive. The sentences are categorized into 3 classes: hate speech, offensive language, or neither in the class column.

4 Methods

We will use a RNN neural network to train our model in order to capture the connection between input words when viewing a sentence as a sequence. The RNN will use a SGD optimizer and a Cross Entropy Loss. We will also include LSTM in the network to get the long-term dependencies in the input.

5 Deliverables

These are ordered by how important they are to the project and how thoroughly you have thought them through. You should be confident that your "must accomplish" deliverables are achievable; one or two should be completed by the time you turn in your Nov 19 progress report.

5.1 Must accomplish

1. Preprocess input sentences to extract text information
2. Train neural network to predict whether a given sentence contains hateful or offensive speeches
3. Generate a list of improper words

5.2 Expect to accomplish

1. Measure hate level of an article
2. Formalize hateful conduct policy
3. Categorize different types of hateful language (racial, sexual, ...)

5.3 Would like to accomplish

1. Apply to other languages
2. Generate non-offensive sentences
3. Take non-text information (such as emojis) into account

References

This section should include citations for: (1) Any papers on related work mentioned in the introduction. (2) Papers describing methods that you used which were not covered in class. (3) Code or libraries you downloaded and used.

Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv. http://arxiv.org/abs/1703.04009

Hate speech and offensive language—Dataset by thomasrdavidson. (n.d.). Data.World. Retrieved October 31, 2022, from https://data.world/thomasrdavidson/hate-speech-and-offensive-language

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv. http://arxiv.org/abs/1301.3781