CS 475/675 Project Proposal

Giovanna Lemos Ribeiro, Isabel Cachola, Kevin He, Olga Karaiman
gribeir3, icachol1, khe7, okaraim1

Abstract

Our motivation is to build a machine learning model that can predict the proportion of negative comments based on the tweet textual data and user information. For our predictive model we will use the tweets' content and use sentiment analysis to analyze the comments.

1 Project choice

Choose either a **methods** or **applications** 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

2 Introduction

The spread of negative content on social media - such as offensive or negative language - is a big problem for our generation. Hate speech in those platforms, for example, have played a crucial role in huge cases of violence across the globe, such as the Rohingya genocide in Myanmar (BBC, The country where Facebook posts whipped up hate 2018), anti-Muslim violence in Sri Lanka (BBC, Sri Lanka vows 'maximum force' against anti-Muslim rioters 2019), and the Pittsburgh synagogue shooting (Robertson, Mele, Tavernise, 11 Killed in Synagogue Massacre; Suspect Charged With 29 Counts 2018). Moreover, social media use has been found to have an impact on the mental health of the population. Its use has been correlated, for example, with depression, low self-esteem and body-image issues in adolescents (Kelly, Zilanawala, Booker, Sacker, 2018), and symptoms of depression and anxiety in the general population during the covid epidemic (Gao et al., 2020). Given the huge potential impact of social media content on the safety and well-being of entire populations, it is important to understand how people interact on those platforms, and how posts elicit negative reactions on the users. A better understanding of what kind of posts are usually followed by comments with low/negative sentiment scores can help us understand how people interact on those platforms, and how different kinds of contents negatively impact the users.

In this project, our goal is to predict the proportion of negative comments on Twitter posts. For this goal, we will use the Twitter API and train a machine learning algorithm for the prediction. The input to the algorithm will be the lowercase tweet and the number of followers of the tweet's author. We will work with more modern approaches for text pre-processing, such as BERT. For more information on the processing of the tweet please see the section on dataset and features. The output of the model will be the predicted proportion of negative comments per tweet. We intend to test different types of machine learning models and select the one with greatest accuracy. We will test linear regression, which has a higher explainability but a more limited hypothesis class, and recurrent neural networks, which has a wider hypothesis class but low explainability.

3 Dataset and Features

Dataset

We will be using Twitter API to extract the data on the tweets (https://developer.twitter.com/en/docs). Moreover, we will use pre-trained models from vader and nltk libraries to extract the sentiment score of the tweets texts and the comments after preprocessing. We will use a total dataset of 100,000 tweets, since this is the daily quota of Twitter API. We will use 15

Features

The features used in the model are:

- Tweet: the full text of the tweet. From this feature we can extract the following features:
  - The content of the tweet (processed using BERT for word embeddings)
  - Sentiment score
  - The number of @ symbols, which represent how many people were tagged
  - The number of hashtags, weighted by their TFIDF importance in twitter
- Author's number of followers: number of followers the author of the tweet has, intended to be used as a normalization factor for the number of retweets
- Number of retweets
- Number of likes
- Proportion of comments with negative sentiment scores - label

Preprocessing

Every tweet text will go through the following preprocessing steps:

- Lower casing
- Elimination of invalid characters or emojis
- Counting of number of mentions (and the elimination of mentions from the text for the following steps)
- Counting of the number hashtags, weighted by their TFIDF importance in twitter (and the elimination of hashtags from the text for the following steps)
- Tokenization using Huggingface library or processing with word embeddings using BERT (we should try both methods)
- Extracting the sentiment score of the tweet using a pre-trained model

Example Data

Text: "This is my first tweet! Hello everyone!! FirstTweet @Friend",
NumFollowers: 20,
NumRetweets: 0 ,
NumLikes: 10,
PercentNegative: 0.1

PercentNegative is only required for training and evaluation, not prediction.

4 Methods

First of all, we will gather 100,000 tweets from the tweet API and create methods to analyse and process the data. We want to understand the distribution of positive and negative comments among all our tweets - and among the specific train/test/validation splits. Moreover, we want to understand if any data (and how much of it) is missing, and the average, mode, and standard deviation of the number of mentions, number of hashtags, retweets, and likes. We will also pre-process the data as detailed above. The data should be organized in a dataframe. We intend to explore some more modern architectures such as word embeddings using BERT to convert the text data into features. BERT is a word embedding method that transforms text into vectors of numbers according to the meaning and context of the words. We want to train different model architectures to predict the proportion of negative comments using the features described in the section above. We plan on creating both a linear regression and a Bidirectional LSTM model. Bi-LSTMs are a graphical model that has a cell state which passes the information through. It is guarded by gates which are affected by a sigmoid layer which determines the amount of information that is allowed to pass through the gate. The hypothesis class for a linear regression is the class of all linear functions. The hypothesis class of Bi-LSTM is much bigger - as a neural network, it is an universal function approximator, given enough data and layers. For our loss function, we will use a mean squared error loss for the linear regression and likely a cross entropy loss for our Bi-LSTMs model, though this may be subject to modification, if we are able to achieve better results with alternatives. Finally for optimization, we will be performing data preprocessing and introducing regularization. For our linear regression model, this will likely come in the form of L2 regularization while for Bi-LSTM, we will be trying out a variety of optimization techniques including but not limited to dropout, early stopping, and varied learning rates.

5 Deliverables

5.1 Must accomplish

1. Using Twitter API, create a dataset and features to be used in the model. We should pre-process our data following the preprocessing steps explained above, and analyse our data quality (is it balanced? is it sparse? Does it have missing values?)

2. Use a pretrained sentiment prediction model to calculate the percentage of negative replies to a tweet
3. Create two models, a linear regression and a Bi-LSTM model, to predict the percentage of negative replies based on the tweet text and author's information

5.2 Expect to accomplish

1. Use more modern neural architectures, like transformers, to make the prediction
2. Try different feature engineering techniques, such as using sentiment based features and using more features related to the hashtags
3. Make an analysis on the model's explainability

5.3 Would like to accomplish

1. Come up with the new applications for this model and discuss those ideas on the final presentation - especially if we can find a way for our results to be applied on twitter
2. Create an API to allow other people to use the model
3. Conduct an analysis of the relationship between user clusters on Twitter and the number of negative replies. The hypothesis is that users who have more cluster overlap receive more negative replies.

References

The country where Facebook posts whipped up hate. (2018, September 12). Retrieved November 04, 2020, from https://www.bbc.com/news/blogs-trending-45449938

Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. International AAAI Conference on Web and Social Media Eleventh International AAAI Conference on Web and Social Media.

Documentation Home | Docs | Twitter Developer. (n.d.). Retrieved November 05, 2020, from https://developer.twitter.com/en/docs

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., . . . Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. Plos One, 15(4). doi:10.1371/journal.pone.0231924

Kelly, Y., Zilanawala, A., Booker, C.,Sacker, A. (2018). Social Media Use and Adolescent Mental Health: Findings From the UK Millennium Cohort Study. EClinicalMedicine, 6, 59-68. doi:10.1016/j.eclinm.2018.12.005

Robertson, C., Mele, C., Tavernise, S. (2018, October 27). 11 Killed in Synagogue Massacre; Suspect Charged With 29 Counts. Retrieved November 04, 2020, from https://www.nytimes.com/2018/10/27/us/active-shooter-pittsburgh-synagogue-shooting.html

Sri Lanka vows 'maximum force' against anti-Muslim rioters. (2019, May 14). Retrieved November 04, 2020, from https://www.bbc.com/news/world-asia-48257299