Introduction

Celebrities suffer from mean comments or trollers on Twitter. A popular TV show Jimmy Kimmel Live has a series of Celebrities Read Mean Tweets. However it is difficult to quantitatively define a troll or mean user in general. To make the problem more specific and solvable, we try to find trollers of a particular celebrity (I choose TS since she is one of my favorites) and try to define what is a troll. Here I use troller, mean user and anti-fan interchangeably because all of them share some same features: (1) Frequently say some swear words to TS (2) Frequently say some bad things (e.g. make some jokes) to TS, not necessarily swear words

The first definition is relatively easy to quantify. I try to identify as many trollers under the first definition as possible, try to find some patterns among them under the second definition, and try to train a classifier not only detect trollers under (1), but detect trollers under (2) as much as possible.

Methods

The idea of the project is as follows:

Getting Data: Obtain a "Gold Standard" dataset which has some mean users and their tweets, as well as fan users and their tweets.

Feature Selection: Extract some feature words which distinguish mean users from fan users, and convert each user's tweets to a numeric vector of frequencies of selected feature words.

Build the model: Train a binary classifier based on the converted data.

However the implement of this idea is not simple, especially getting the "Gold Standard" dataset. Tagging a user as troll or fan is manually done in this project. The details involved in obtaining the data are described in Data Collection section. Then we present some EDA results. Feature words are manually selected from high frequency words, in order to incorproate words that could distinguish trolls from fans. We introduce the details of feature selection and the statistical model in the following section.

Data Collection

We use R package "TwitteR" in this project. The search function in this package use a Twitter API.

Preliminary steps

Under the first definition of trolls, we need to find users that tweet swear words and mention Taylor Swift. So we first come up with a badword dictionary, which contains 172 swear words and their internet variations. For each swear word in the dictionary, we search for 300 tweets that contain the swear word and mention Taylor Swift (the returned results are all less than 300). Then we get the screen usernames of those returned tweets as trolls. Similarily we search for 3000 tweets that contain "love" and mention TS, and get the usernames as fans. Since there are much more fans than trolls, we just use a single keyword "love" to identify fans. However we find this is not enough to correctly identify trolls and fans. So we manually inspect each tweets and decide whether it is a true troll/fan tweet. We delete tweets that are not troll in the troll set and tweets that are not fan from the fan set. Then we get the unique screen usernames and identify them as true trolls and fans.

Raw Data

We manually identify 291 trolls and 603 fans. We search for most recent 200 tweets for each individual. The search is splited to several search requests since the API rate limit is reached by a single search. We collect 26,205 tweets from those 291 mean users, as well as 41,847 tweets from those 603 fan users.

Data Cleaning

Tweets contain URL link, hex encoded emoji and other characters that we do not want to include in our analysis. In order to do text mining, we write a MyClean function to get a clean text of all tweets.

Also for sentimental analysis and feature selection, we intgrate tweets from each person as a single text.

Exploratory Data Analysis

Wordcloud

Calculate word frenqucy of all tweets from trolls and fans seperately. Select top 200 words of trolls and fans. Manually delete meaningless words and words of high frenqucy in both sets to better see difference in word frenquency. Words in troll set includes fan set:

Sentimental Analysis

sentiment for each tweet

sentiment for each person

Statistical Modeling

Reproducibility

We manually tag a user as troll or fan, and manually select feature words. These 2 steps are not reproducible. But given a "Gold Standard" dataset and fix the feature words, other parts of the analysis are reproducible.

Result

Conclusions

References