Assessing BERT's Ability to Judge Moral Scenarios
Using Reddit's "Am I The Asshole" as a Database

Exposé

Author: Scott Fletcher
fletches@hu-berlin.de

Supervisors: Ulf Leser
Anke Lüdeling

Introduction

Sentiment Analysis is a field of Natural Language Processing (NLP) which focuses on extracting subjective opinions from natural language texts. It has been used to gauge public opinion on brands [1], assess internet marketing possibilities [2], and analyze app user-reviews to gauge which features are liked/disliked [3] to name a few examples. This project will use BERT [4], a new technology from Google which is currently setting records in Sentiment Analysis, to predict and mimic human moral judgment using the Reddit community http://www.reddit.com/r/amitheasshole as a dataset.

Motivation

The growing popularity of internet forums offers us a snapshot of sociolinguistic relationships across a broad range of topics. Information shared in this way has been difficult to extract systematically due in large part to its informal and extremely varied presentation. New technologies such as BERT [4] use enormous training sets of unlabelled information to systematize linguistic context, supplementing existing Machine Learning techniques to be able to parse natural text more capably. [5]

Due to this, the scope of information available through systematically analyzing written text has grown immensely. The reams of text available through the internet contain enough data to systematically analyze linguistic nuances that have up until now been considered exclusively perceptible to humans. The difficulty in analyzing these behaviors has been finding enough labelled examples of texts exhibiting said behaviors.

The subreddit r/AmITheAsshole is an online community centered around collecting and labelling natural language text. The criteria are the morals of the user-base; whether the users believe a person in a given text acts justly or not. The result is what one may

consider a ‘morality handbook,’ consisting of well over 100,000 unique everyday stories complete with social commentary. The question this project attempts to answer is, can new technologies in the field of NLP help computers learn this handbook in the same way a human can?

What is BERT

BERT or Bidirectional Encoder Representations from Transformers [4] is a context oriented language representation model. It builds upon the ELMo [6] model of word embeddings, which considers the entire sentence around a given word before assigning it an embedding. BERT was trained using large unmarked corpuses like wikipedia. It currently is setting records for predictive language and context interpretation tasks.

What is Reddit

Reddit is an internet forum which allows anonymous users to post, comment on, and rate submissions and comments. A submission consists of any combination of text, links, images, or videos and belongs to one of many **Subreddits** on the site. A comment consists of solely text or links and is attached to either a submission or another comment. Users rate submissions and comments using the **Upvote / Downvote** system. A **Subreddit** is a forum focused around a single topic or theme which users can subscribe to. According to Reddit, there are currently over 330 million unique Reddit accounts and over 138 thousand active Subreddits.

Reddit is among the largest English-speaking internet communities and its easy accessibility and anonymity allow for viral [read: information-rich] conversations to form around taboo topics such as morally unacceptable scenarios. [7]

What is r/AmITheAsshole

r/AmITheAsshole is a Subreddit in which users can post first-person narrative stories and, as the name suggests, ask the Reddit user-base whether they are the ‘asshole’ [bad guy] in the scenario. Users can then cast one of 4 different votes in the form of a comment:

NTA: Not the A-hole
The poster of the story is not the bad guy, but someone else in the story is (can be specified in the comment)

YTA: You’re the A-hole
The poster of the story is the bad guy and the rest of the people involved are not.

NAH: No A-holes Here
Nobody in the story is a bad guy.

ESH: Everyone Sucks Here
Everyone in the story is a bad guy; no one acted reasonably.

Comments can then be upvoted by users who agree with the vote and argumentation, and at the end of a 24 hour period, the

'verdict' is automatically selected from the highest rated comment.

Goal

We will approach the task by using the submission text to predict the response of the r/AmITheAsshole community. This will be done using BERT's native classification and Support Vector Machine (SVM) classification as a baseline. The success rate will be determined by 10-fold cross-validation. The goal is to achieve a success rate significantly greater than random classification.

Related Works

Reddit has been a platform of scientific research since its founding in 2005. Choi et al. [8] research the characteristics of threaded conversations on Reddit as well as the factors which lead to viral threads. De Choudhury and De [7] research the effect of anonymity on willingness to disclose issues relating to mental illness on Reddit and further establish a connection between Reddit's 'throwaway account' culture and discussing taboo topics.

NLP studies based on Reddit have been able to find complex meta-information in varying contexts. Tigunova et al. [9] use Subreddits IAMA and AskReddit to train a Hidden Attribute Model to recognize characteristics of speakers (ie. professions, hobbies, personal relationships, etc.). Harrigian [10] uses Named Entity Recognition (NER) to associate over 65,000 reddit accounts with their geographic location.

Reddit is especially useful in analyzing typically taboo topics and behavior. Greaves and Dykeman [11] use Reddit submissions to systematically recognize linguistic cues indicative of Non-Suicidal Self-Injury. They found correlations between certain grammatical structures (over-use of the first person singular, under-use of the first person plural) and certain keywords to the categorization of the texts. Sekulić et al. [12] use linguistic cues put forth by psychological research to analyze users' Reddit post history and predict bipolar disorder. Schrading et al. [13] test a variety of linear regression models in training a system to recognize Reddit posts dealing with domestic abuse. They achieved an F1-score of 0.863 with a Random Forest model.

Reddit is also one of the few mainstream social media outlets where morally ostracized groups can be researched in large scale. Mittos et al. [14] use NLP and computer vision to examine how the far right use popular genealogy tests to back neo-nazi rhetoric on specific Reddit forums. Finkelstein et al. [15] scrape the radical right-wing websites 4chan (/pol/) and Gab for antisemitic posts and conduct a diachronic study on the spread of these posts to other popular websites, including Reddit.

Method

There are four steps in this project:

1. data retrieval,
2. data analysis,
3. experimentation, and
4. result analysis

Data retrieval

The texts used in this project are all publically available via www.reddit.com/r/amitheasshole. I will use two APIs in order to extract the necessary information; Pushshift.io [16], a third party web crawler specifically designed around Reddit's architecture, and Reddit's own PRAW python API library [17].

I will extract submissions from the last 3 years which have at least 20 comments, each comment representing a unique vote. For each submission, I will extract up to 100 of the top rated vote comments.

Comment and submission information will not be published. Data stored on my computer and hard drive will be encrypted and password protected. The data is solely intended for research purposes.

Data labelling

Submission texts will be labeled according to the votes in their comments. Each comment may contain one of 4 relevant votes:

Positive	Negative
NTA	YTA
NAH	ESH

In comments containing more than one vote, only the first will count. Comments without votes or containing one of the two irrelevant votes (SHP, INFO) are ignored.

Each comment has an upvote / downvote score. For each of the 4 votes, the sum of the scores of all comments containing it will be calculated. These 4 sums will form the label of the submission.

For example, say a submission has 10 comments. We look at the vote and the score of each comment:

Vote	Score
NTA	1000
NTA	600
YTA	500
NTA	200
NAH	150
ESH	100
[none]	200
YTA	50
INFO	10
SHP	5

The resulting label is:

NTA	NAH	YTA	ESH
1800	150	550	100

Experimentation

Two experiments will be conducted:

1) 4-class classification (NTA, NAH, YTA, ESH)
2) 2-class classification (positive, negative)

Experiments will be conducted using 10-fold cross-validation. This involves splitting the dataset into 10 equal-sized parts. This will be done using stratified sampling to ensure that each part has a roughly even distribution of classes. Each experiment will be conducted 10 times using 9 of the parts as a training set and the remaining 1 part for testing. In this way, each individual submission is used for testing exactly once.

Experiments will use BERT [4] to parse and classify the texts. A baseline will be calculated for each experiment using SVM classification.

Result Analysis

The amount of True Positives, False Positives, True Negatives, and False Negatives [TP, FP, TN, FN] will be recorded for each test. Afterwards, Recall and Precision as well as an F1-Score (the harmonic mean of precision and recall) will be calculated.

Bibliography

[1] Arora, D., Li, K. F., & Neville, S. W. (2015, March). Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications* (pp. 680-686). IEEE.

[2] Melville, P., Gryc, W., & Lawrence, R. D. (2009, June). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1275-1284). ACM.

[3] Guzman, E., & Maalej, W. (2014, August). How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)* (pp. 153-162). IEEE.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. Pdf*.

[6] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[7] De Choudhury, M., & De, S. (2014, May). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[8] Choi, D., Han, J., Chung, T., Ahn, Y. Y., Chun, B. G., & Kwon, T. T. (2015, November). Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks* (pp. 233-243). ACM.

[9] Tigunova, A., Yates, A., Mirza, P., & Weikum, G. (2019). Listening between the Lines: Learning Personal Attributes from Conversations. *arXiv preprint arXiv:1904.10887*.

[10] Harrigian, K. (2018). Geocoding without geotags: A text-based approach for reddit. *arXiv preprint arXiv:1810.03067*.

[11] Greaves, M. M., & Dykeman, C. (2019). A Corpus Linguistic Analysis of Public Reddit Blog Posts on Non-Suicidal Self-Injury. *arXiv preprint arXiv:1902.06689*.

[12] Sekulić, I., Gjurković, M., & Šnajder, J. (2018). Not Just Depressed: Bipolar Disorder Prediction on Reddit. *arXiv preprint arXiv:1811.04655*.

[13] Schrading, N., Alm, C. O., Ptucha, R., & Homan, C. (2015). An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2577-2583).

[14] Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2019). " And We Will Fight For Our Race!'" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. *arXiv preprint arXiv:1901.09735*.

[15] Finkelstein, J., Zannettou, S., Bradlyn, B., & Blackburn, J. (2018). A quantitative approach to understanding online antisemitism. *arXiv preprint arXiv:1809.01644*.

[16] http://pushshift.io

[17] http://praw.readthedocs.io