Reg Gonzalez
CS 4395.001

ACL Paper Summary

Introduction:

The title of the paper I chose is “Ruddit: Norms of Offensiveness for English Reddit Comments”. The authors of this paper are Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. The affiliations are ILLC, University of Amsterdam (Hada, Sudhir, and Shutova); Facebook AI, London (Mishra); Department of Informatics, King’s College London (Yannakoudakis); and National Research Council Canada (Mohammad).

Problem:

The problem the authors wanted to address in this paper was to be able to detect offensive language on social media platforms (specifically Reddit in this case) in better ways than the current automatic methods. Previously, methods used to detect offensive language very much relied on categorical labels in datasets (e.g., labeling data as “hate speech”, “abusive language”, “spam”, etc.). However, many times these categories overlap with each other, which creates ill-defined boundaries about what may or may not be considered “offensive language.”

Along with that, in real life there is also a spectrum of offensiveness. Language just isn’t a binary “yes, that’s offensive” or “no, that’s not offensive”. Instead, there is a scale of what can be considered offensive and what things are considered *more* offensive than others.

Prior work:

When it comes to looking at the prior work for this specific research, the paper highlights two important aspects: offensive language datasets and best-worst scaling (BWS). We’ll first take a look at the offensive language datasets.

When it comes to the datasets, the paper sites several datasets used to detect offensive language, including some by researchers Zeerak Waseem and Dirk Hovy (2016), Davidson et al. (2017), and Founta et al. (2018). These particular offensive language datasets were created using data from Twitter. Waseem and Hovy’s dataset found/used terms that frequently occurred in offensive tweets. The dataset created by Davidson et al. instead used a list composed to “hate-related” terms in order to extract tweets that they deemed offensive. The dataset created by Founta et al. made use of sentiment analysis by checking for the occurrence of offensive terms. In many of these datasets (as well as many other datasets), “offensive” language was divided into two groups: explicitly offensive and implicitly offensive. Explicitly offensive language contained profanity and swear words while implicitly offensive language didn’t. The paper goes on to discuss other pre-existing offensive language datasets, many of which use categorical labels to detail what type of language is considered offensive or not.

The second thing the paper discusses regarding prior work is best-worst scaling (BWS). Best-worst scaling was proposed by Jordan J. Louviere in the lates 1980s/early 1990s. The purpose of BWS is to produce more reliable fine-grained scores then simply relying on rating scales. In all, BWS is a methodology used in surveys that asks a user to pick which item from an n-tuple is the “best” and which is the “worst”. Here, “best” and “worst” are relative to whatever it is you’re interested in researching. Usually in this n-tuple, n = 4; however, that number can vary. When it comes to its application in natural language processing, this type of methodology can be used to create datasets including word-sense disambiguation, word/phrase sentiment analysis, emotional intensity, etc. For this paper, BWS is one technique used to calculate offensiveness scores.

Unique contributions of this paper:

The first unique contribution made in this paper was the authors introducing their own dataset to work with. This dataset was composed of 6,000 English Reddit comments with real-valued, fine-grained scores ranging from -1 to 1. A value of -1 represented a comment that was least offensive, while a value of 1 represented a comment that was most offensive.

This paper also made use of the comparative annotation setup. Using this setup, annotators are given two or more comments and are asked what is the most offensive and what is the least offensive. According to the paper, using this kind of setup is better than traditional rating scales because it mitigates annotation biases and categorical labels (since the annotators are not asked to label the data). Finally, it also alleviates annotator desensitization (i.e., an annotator might be inclined to rate something as more or less offensive when comparing it to their previous annotation).

The authors also made note of how offensive language and behavior is often linked with strong emotions. In order to capture this idea in the research, they decided to sample comments from the NRC VAD lexicon (a lexicon of 20,000 English words) that were measured based on three dimensions of emotions: valence, arousal, and domiance. Valence refers to a positive or negative feeling, arousal refers to an excited or calm feeling, and dominance refers to a powerful or weak feeling. These words were measures to have real-valued scores based on those three dimensions, ranging from 0 to 1. Much of their sampling came from comments with low valence and high arousal values.

When it comes to how they sampled the comments, they decided to use a hybrid approach. That is, they took their comments from three different categories: topics from specific subreddits, randomly sampled comments from Reddit, and comments from the subreddit r/ChangeMyView (a subreddit that has an inherently large argumentative property). The comment types within each category is evenly split amongst three types: randomly sampled, comments with low valence, and comments with high arousal.

In regards to the annotation, annotations were made using Amazon Mechanical Turk. According to the paper, in order to ensure a minimized effect on the annotators' mental health, annotators were given a limit on how many annotations they could make. The authors themselves annotated 5% of the dataset and used those as what they called "gold questions". They said that if annotators had annotations that were below 70% on those gold questions, their work would be discarded. The annotators also made use of best-worst scaling. They were shown four comments and were asked to chose which one was the most offensive and which one was the least offensive. Using these annotations, they were then able to calculate offensiveness scores (which were the percentage of times the comment was picked as most offensive minus the percentage of times the comment was picked as the least offensive).

In order to make sure that these annotations were reliable, the authors used a metric called the Split-Half Reliability metric. Here, they took their annotated data, randomly divided them into two even groups, and created ranking lists based on each group. Those two ranking lists were then compared—using some correlation metric, r. This was then repeated 100 times, were r was averaged out in the end. A high correlation metric indicates a high reliability.

Authors' evaluation of their work:

The authors were able see that their dataset of comments ranged most supportive/least offensive to least supportive/most offensive. They were able essentially categorize them into one of five groups: mostly supportive, supportive/neutral, neutral/potentially offensive, offensive (including many implicitly offensive comments), and very offensive (mostly explicitly offensive).

In regards to computational modeling, the goal was to predict the offensiveness score for a given comment. In order to do this, three different models were used: BiLSTM, BERT, and HateBERT. With BERT and HateBERT specifically, those needed to be fine-tuned in order to get more accurate results. Across all three models, they performed 5-fold cross validation.

To get a better idea of what exactly the models were learning, the authors needed to create a few dataset variations. The first of which was called Ruddit, which was just the complete dataset. The second was called identity-agnostic, which was to aid in seeing the effective of identity terms within offensive comments. The third was the no-swearing dataset, which as its name suggests, was to aid in seeing the effect of disregarding profanity within offensive comments. Finally, the last dataset variation was the reduced-range set, which consisted of comments that had an offensiveness score ranging from -0.5 to 0.5

When it came to comparing the Ruddit and identity-agnostic datasets, the HateBERT model outperformed both BERT and BiLSTM. It was also noted that HateBERT performed slightly better on the Ruddit dataset than the identity-agnostic dataset. Comparing the no-swearing and reduced-range datasets HateBERT once again outperformed BERT and BiLSTM. HateBERT also performed better on the no-swearing dataset than it did the reduced-range set.

Number of citations per author (based on Google Scholar):

Rishav Hada - 3 articles, cited by 16
Sohi Sudhir - 3 articles, cited by 10
Pushkar Mishra - 19 articles, cited by 401
Helen Yannakoudakis - 59 articles, cited by 1913

Saif M. Mohammad - 127 articles, cited by 17026
Ekaterina Shutova - 103 articles, cited by 3042

The author with the most citations in this group is Saif M. Mohammad.

My thoughts on the importance of the authors' work:

I think the work conducted and shown by the authors of this paper is important for a number of reasons. Firstly, I think it's just interesting to see what types of comments are considered offensive, the spectrum of offensiveness of comments, and how we can automatically detect what comments are on that spectrum. And considering that language is consistently evolving and changing, I think that from a research perspective, it'll be interesting to look at this dataset and the research that was conducted/presented here and compare it to language years from now.

When it comes to NLP specifically, I think that looking at the supportiveness/offensiveness of comments could come in handy when programming tools such as chatbots. This way, chatbots and other intelligent systems may be able to figure out if a comment is supportive, offensive, or neutral.

Finally, perhaps the most prevalent place offensive comments can be found nowadays is through social media. There have been many discussions regarding social media and how it's played a role in shaping our mental health. The use of offensive comments and language is at the forefront of that discussion and by being able to detect these types of comments automatically, we can perhaps do a better job of limiting them. Not only would it mitigate some of the effects that negatively impact people's mental health, but it could also make things like healthier discussion forums and chat rooms.