[Proposal]: Identifying and Categorizing Offensive Language on Twitter

Arunav Saikia
M.S. Data Science
Indiana University - Bloomington
arsaikia@iu.edu

Abstract

Offensive language, hate speech and cyberbullying have become increasing more pervasive in social media. Individuals frequently take advantage of the perceived anonymity on social media platforms, to engage in brash and disrespectful behaviour that many of them would not consider in real life. The goal of this project is to use a hierarchical model to not only identify tweets/messages with offensive language but categorize the type and the target of offensive messages on social media.

1 INTRODUCTION

Methods and techniques for abuse detection are important to detect cyberbullying and hate speech on social media platforms, which affect users from all demographics. It is estimated that approximately 680 million tweets, 4.3 billion messages on Facebook and 300 billion emails are sent everyday (1). A survey by Pew Research Centre states that 40% of adults have personally experienced harassment online (2). There has also been an increase in instances of hate posts that stir violence across societal groups in multiple countries (3). Hate posts with aggressive content if left unmoderated can polarize communities to become dangerously violent and can even lead to riots. White supremacist attacks in the US (4), ethnic cleansing of Rohingya Muslim minorities in Myanmar (5), mob lynching and communal violence in India (6) are just a few examples of this mayhem in recent years.

While most social media websites have inbuilt provisions for its users to flag and report offensive or harmful content, only 8.8% victims actually use this option (2). This statistic highlights the importance of automatic methods to curb and moderate offensive content as well as abusive profiles because passive and manual methods are neither effective nor scalable in real time.

Most previous work on this topic focused on building supervised classification system to detect specific types of offensive contents - eg racist/sexist/neither (7) or hateful/offensive/neither (8) . In this work, we propose an end-to-end framework to identify the type and target of offensive content on social media website - Twitter. The assumption is that the target of offensive messages is an important indicator that allows us to discriminate between, e.g., hate speech, which often consists of insults targeted toward a group, and cyberbullying, which typically targets individuals etc. This framework will enable organizations to automatically detect, categorize and take targeted measures based on the type of offensive language.

2 RELATED WORKS

Using supervised learning to detect aggressive or abusive language from text is a widely used approach. Early approaches used simple heuristic-based features to represent text. Spertus (9) used handcrafted lexical features to capture syntax and semantics of text, followed by a decision tree classifier to detect abusive messages. Consequently, Yin et al (10) came up with a formal definition for 'harassment' on Web2.0 and used bag-of-words, TF-IDF and heuristic based semantic features to detect harassment in online forums.

Advances in neural networks has led to vector-based representation of words. These low dimensional vectors ensure that words having similar context occur together in the embedding space. These word embedding techniques like word2vec (11) and GloVe (12) led to classifiers (13) leveraging low dimensional vector representation in addition to linguistic (length, number of punctuations, etc) and syntactic features (POS) applied to comments on social media. Pavlopoulos et al (14) suggested using RNN and CNN based architectures with word

Tweet	A	B	C	Category
@USER He is so generous with his offers.	NOT	—	—	Clean
WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—	Profane
@USER she is human garbage that is what is wrong	OFF	TIN	IND	Cyber-bullying
@USER Leftists showing yet again how vile they can be	OFF	TIN	GRP	Hate speech

Table 1: Four tweets from the OLID dataset, with their labels for each level and final categorization.

embeddings to detect abusive user comments on Wikipedia.

In (15) Singh et al use CNNs and LSTMs to detect aggression in social media text involving multilingual speakers. Detecting and filtering such content on social media is a challenging task because most multilingual populations use 'code-mixing' to express their opinions. Code mixing refers to the practice of using linguistic units from multiple languages in a single sentence. The authors in (15; 16) focus their work on detecting aggressive markers on code mixed Hindi-English text generated on Facebook and Twitter by Indians.

All of the above approaches solely rely on linguistics and NLP to detect abusive content. But recent advances in graph mining and graph representation learning has led to social network analyses approach to tackle this problem. Emphasizing the principle of homophily, Mishra at el (17) mentioned that abusive content comes from users who share a set of common stereotypes and form communities around them. The authors proposed a representation learning approach to create 'profiles' of tweeter users. Specifically, the authors leveraged a node embedding framework called node2vec (18) to learn node representations from a community graph where nodes represent tweeter users and edges represent follower-following relationship. The authors used these profile representations along with text-based features from (20) and a GBDT classifier to detect racist and sexist tweets.

Most of these work do not distinguish between the different types of offensive language that exists in the social media platforms. They either 1) use terms like 'cyberbullying', 'abusive', 'hate speech' interchangeably when there exist clear definition (21) for each form of offensive language based on the content and the target or 2) work on specific forms of offensive language like racist vs sexist (7) while ignoring others. Identification and proper categorization of all forms of offensive language is important so that the organizations and government entities can take proper well informed mitigation steps. Our proposed method is a hierarchical approach with three sub-tasks:

- Sub-task A: Offensive language identification as offensive (OFF) or not offensive (NOT)
- Sub-task B: Automatic categorization of offense types as either targeted insult (TIN) or untargeted (UNT)
- Sub-task C: Offense target identification as individual (IND), group (GRP) or other (OTH)

Our end-to-end hierarchical framework allows us to use the combination of labels in each sub-task to identify specific types of offensive language like profane or hate speech or cyberbullying etc as shown in Table 1

3 DATA

The data used in this work was released as part of *SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media* (OffensEval) (22). Offensive Language Identification Dataset (OLID) dataset, is a large collection of English tweets annotated using a hierarchical three-layer annotation model to distinguish between whether the language is offensive or not (A), its type (B), and its target (C). It contains 14,100 annotated tweets divided into a training partition of 13,240 tweets and a testing partition of 860 tweets. Each level is described in more detail in the following subsections.

3.1 Level A:

- **Not Offensive (NOT):** Posts that do not contain offense or profanity;
- **Offensive (OFF):** Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words

A	B	C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	-	524	27	551
NOT	-	-	8,840	620	9,460
All			13,240	860	14,100

Table 2: Distribution of label combinations in OLID.

3.2 Level B:

- **Targeted Insult (TIN):** Posts containing insult/threat to an individual, a group, or others;
- **Untargeted (UNT):** Posts containing non-targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

3.3 Level C:

- **Individual (IND):** Posts targeting an individual. This can be a famous person, a named individual or an unnamed participant in the conversation. Insults and threats targeted at individuals are often defined as cyberbulling.
- **Group (GRP):** Posts targeting a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristic. Many of the insults and threats targeted at a group correspond to what is commonly understood as hate speech.
- **Other (OTH):** The target of these offensive posts does not belong to any of the previous two categories (e.g., an organization, a situation, an event, or an issue).

The distribution of the labels in OLID is shown in Table 2. The dataset was annotated using the crowdsourcing platform Figure Eight. More information about the data collection and data annotation process can be found here (23). A few high level statistics for the tweets in Sub-task A can be seen in Figure 1, Figure 2 and Figure 3.

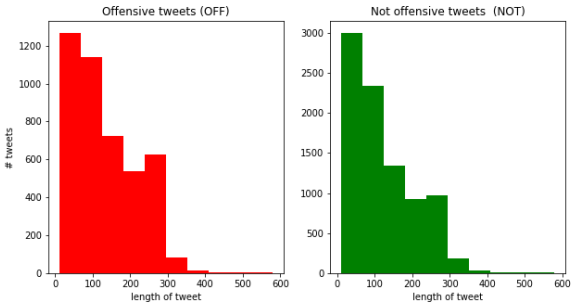

Figure 1: Histogram of # characters in tweets for sub-task A

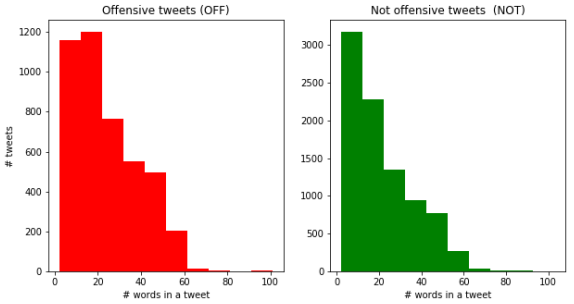

Figure 2: Histogram of # words in a tweet from sub-task A

4 PROPOSED METHOD

5 RESULTS

6 CONCLUSION

7 SOURCE CODE

Source code for all experiments and results can be found here `https://github.com/arunavsk/OffenseEval2019`.

8 ACKNOWLEDGEMENTS

The author wishes to thank Prof Allen Riddell for his instructions and support. This work was part of Social Media Mining (ILS-Z639) course for Fall 2020 at Indiana University, Bloomington.

References

[1] Online - `https://www.gwava.com/blog/internet-data-created-daily`

[2] Online - `https://www.pewresearch.org/internet/2017/07/11/online-harassment/`

[3] Online - `https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons`

[4] Online - `https://www.newyorker.com/magazine/2017/02/06/inside-the-trial-of-dylann-roof`

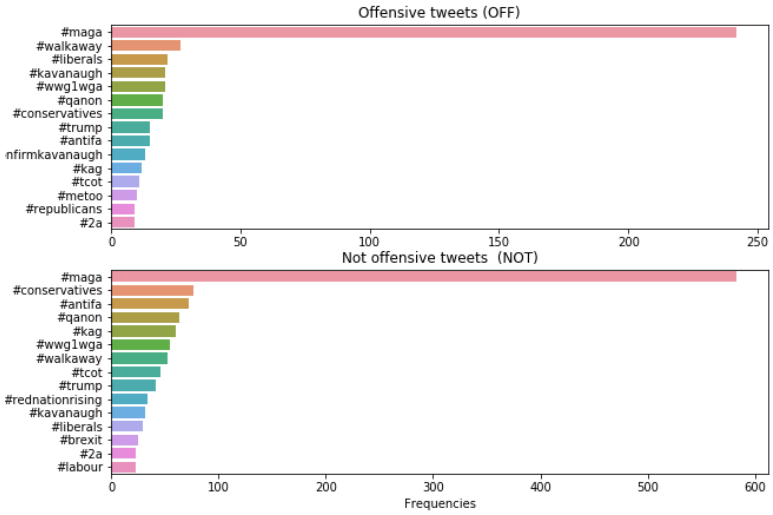

Figure 3: Histogram of hashtags in sub-task A

[5] Online - https://iwpr.net/global-voices/how-social-media-spurred-myanmars-latest

[6] Online - https://www.washingtonpost.com/graphics/2018/world/reports-of-hate-crime-cases-have-spiked/-in-india/

[7] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.

[8] Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. "Automated hate speech detection and the problem of offensive language." arXiv preprint arXiv:1703.04009 (2017).

[9] Spertus, Ellen. "Smokey: Automatic recognition of hostile messages." In Aaai/iaai, pp. 1058-1065. 1997.

[10] Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2 (2009): 1-7.

[11] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)

[12] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[13] Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. "Abusive language detection in online user content." In Proceedings of the 25th international conference on world wide web, pp. 145-153. 2016.

[14] Pavlopoulos, John, Prodromos Malakasiotis, and Ion Androutsopoulos. "Deeper attention to abusive user content moderation." In Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 1125-1135. 2017.

[15] Singh, Vinay, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava. "Aggression detection on social media text using deep neural networks." In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 43-50. 2018.

[16] Bohra, Aditya, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. "A dataset of Hindi-English code-mixed social media text for hate speech detection." In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pp. 36-41. 2018.

[17] Mishra, Pushkar, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. "Author profiling for abuse detection." In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1088-1098. 2018.

[18] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855-864. 2016.

[19] Mishra, Pushkar, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. "Abusive language detection with graph convolutional networks." arXiv preprint arXiv:1904.04073 (2019).

[20] Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma. "Deep learning for hate speech detection in tweets." In Proceedings of the 26th

International Conference on World Wide Web Companion, pp. 759-760. 2017

[21] Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51, no. 4 (2018): 1-30.

[22] Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)." arXiv preprint arXiv:1903.08983 (2019).

[23] Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "Predicting the type and target of offensive posts in social media." arXiv preprint arXiv:1902.09666 (2019).

[24] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710. 2014.

[25] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[26] Hamilton, Will, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." In Advances in neural information processing systems, pp. 1024-1034. 2017.

[27] Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).

[28] Schlichtkrull, Michael, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. "Modeling relational data with graph convolutional networks." In European Semantic Web Conference, pp. 593-607. Springer, Cham, 2018.

[29] Zhang, Chuxu, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. "Heterogeneous graph neural network." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 793-803. 2019.

[30] Wang, Xiao, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. "Heterogeneous graph attention network." In The World Wide Web Conference, pp. 2022-2032. 2019.