24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

The Refugee Experience Online: Surfacing Positivity Amidst Hate

Shriphani Palakodety[12] and **Ashiqur R. KhudaBukhsh**[13] and **Jaime G. Carbonell**[4]

Abstract. How can Artificial Intelligence help a stateless minority from online abuse? Research efforts in hate speech detection thus far have largely focused on identifying and subsequently filtering out negative content that specifically targets them. In this paper, we highlight a recent work [8] which tackles a different aspect of web-vulnerability of marginalized communities: sparsity of pro-minority voices championing their cause. The highlighted paper advocates that blocking hate alone may not be sufficient in these cases as the internet shapes community perception to a great extent in modern times and supportive comments to a vulnerable community serve a different purpose. Using an Active Sampling approach, the paper constructs a nuanced *voice-for-the-voiceless classifier* that automatically discovers comments supporting a (allegedly) persecuted minority. In the context of the Rohingya refugee crisis, one of the biggest humanitarian crises of modern times, the paper presents promising results that can substantially aid content moderation efforts in finding positive content supporting the Rohingyas.

1 INTRODUCTION

How can Artificial Intelligence help a stateless minority from online abuse?

Hate speech detection is a widely-studied research challenge [10] that seeks to detect communication disparaging a person or a group on the basis of race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [5]. Online attacks against at-risk marginalized communities can be harmful in two different ways. First, exposure to hate speech may play a role in increased prejudice among the bystanders through desensitization [12]. Second, the attacks may not remain confined to the online world; evidence of a close, causal link between online hate speech and offline violence targeting refugees has been reported [4].

A typical strategy to counter online hate speech is detection and subsequent web-moderation (e.g., flagging a user, deleting offensive posts). However, in the context of helping a marginalized community, blocking the hateful content may not be enough; it would be particularly beneficial if content supportive to the minorities is also emphasized in parallel. The highlighted work [8] presents a new outlook to address the challenge of improving the online refugee experience through promoting *help speech*, i.e., content that champions a marginalized community. The paper emphasizes a different aspect of web-vulnerabilities of the refugees: sparsity of pro-minority voices supporting their cause and raises an important point that in a discussion largely disparaging to a minority, the minority may not be able to defend themselves because of their absence stemming from (1) unfamiliarity with a global language (2) limited internet access and (3) most importantly, immediate physical safety being their highest priority. To this end, in the context of the Rohingya refugee crisis, [8] presents a *voice-for-the-voiceless* classifier that automatically discovers content favoring the Rohingyas. Specific to the Rohingya refugee crisis [3], *voice-for-the-voiceless* is defined as content that urges organizations (such as UN) or common people to help them, expresses empathy or condemns the (alleged) oppressor or advocates for their rights (a detailed definition is presented in [8]).

2 WEB-VULNERABILITY

The highlighted paper constructed a corpus of YouTube comments relevant to the refugee crisis. Count-based statistics and sentiment analysis revealed that a large fraction of comments in the data set were disparaging to the Rohingyas with several comments equating them to terrorists. Resettlement of the Rohingyas was a highly debated issue. One particular alarming finding was *hell* was one of the top completions for the text-template `[send them to]`.

3 RESEARCH CHALLENGES

Rare positives: The goal of the *voice-for-the-voiceless* classifier is to detect pro-minority voices drowned out by a large number of hateful or neutral comments targeting them. Essentially, this is a learning problem with class imbalance where positives are rare. However, in order to construct an effective *voice-for-the-voiceless* classifier, a balanced data set with considerable number of positives and negatives is required. Moreover, due to the nuanced definition of the target concept, examples covering all sub-categories of positives (and negatives) in the training set can improve performance in the wild.

Linguistic challenges: The events surrounding the Rohingya crisis occurred in South and South-East Asia. The bulk of internet users in this part of the world does not speak English as their first language. As a consequence, the corpus features significant spelling and grammar disfluencies. A few examples are provided in Table 1. Adding to the complexity, the region is highly linguistically diverse with a substantial mix of languages. Thus the challenge is two-fold - (i) discover and surface content using methods that can handle short, noisy text, and (ii) handle the large linguistic diversity.

The highlighted paper leverages modern NLP developments to address both these challenges. First, comment-embeddings - distinct real valued vectors for each comment - are obtained using [6] and the sub-word modeling allows for robust representations even for comments with high levels of misspelled words. Next, a technique from

[1] Ashiqur R. KhudaBukhsh and Shriphani Palakodety are equal contribution first authors.
[2] Onai, USA, email: spalakod@onai.com
[3] Carnegie Mellon University, USA, email: akhudabu@cs.cmu.edu
[4] Carnegie Mellon University, USA, email:jgc@cs.cmu.edu

[7] is used that builds a minimally supervised language-identification system to extract out the English comments.

it is a bit **hearting** that all the developed countries are **sailfish**
...**vi** want myanmar army goverment to the icc **kireminal courd** justice and **vi** want full **setizenthip**...

Table 1: Examples of the spelling and grammar disfluencies observed in the corpus.

4 ACTIVE SAMPLING

In a similar spirit as [2], [8] melds recent advances in sentence embeddings [6] with Active Learning literature and presents an Active Sampling approach to expand a seed set of example comments exhibiting *voice-for-the-voiceless*. Using a seed set that could be unambiguously annotated, and a set of manually written comments i.e. not already present in the corpus but constructed by the annotators to illustrate the types of positive comments, nearest-neighbor sampling in the comment-embedding space was conducted to obtain a diverse set of positive and negative examples. Manual inspection reveals that the similarity methods are able to utilize the annotator-written illustrative comments and retrieve similar comments from the corpus. As mentioned, the sub-word component of the embeddings mitigates potential issues arising from spelling variations and misspellings.

Sampling in the comment-similarity space is then combined with rounds of one-sided certainty sampling [11, 1], and traditional uncertainty sampling. The end-result of this sequence yields a balanced data-set containing nearly equal numbers of positive and negative class examples. The method shows promise in future text-classification tasks with drastic class imbalance.

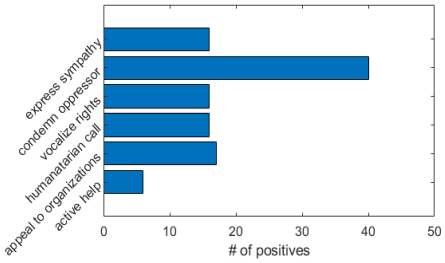

Figure 1: Breakdown of positive comments found in the wild. A single comment can satisfy multiple criteria (Figure taken from [8]).

5 CLASSIFICATION

The highlighted paper used a simple SVM baseline with n-gram features ($n \in [1, 3]$) and obtained a performance of precision: 73.65, recall: 79.39, F1: 76.34, and accuracy: 75.38. An across-the-board improvement was reported when the comment-embeddings discussed above were included alongside the n-gram features (precision: 76.49, recall: 80.30, F1: 78.28, and accuracy: 77.71). This is attributed to the embeddings' ability to perform well even amidst misspellings.

6 PERFORMANCE IN THE WILD

The highlighted paper deployed the classifier on a set of unseen comments - i.e. they were not part of the train or test sets. Evaluation reveals a substantially higher level of positive comments (88%) retrieved compared to mere random sampling (10.67%). Further annotation of the classifier-discovered comments into various sub-categories reveals a highly diverse mix of comments spanning all sub-categories presented in the definition (shown in Figure 1). The sampling strategy's merits are thus established in (i) balanced data-set construction, (ii) diversity of uncovered positives.

7 CONCLUSION

The highlighted paper discusses methods and a new task in a vital humanitarian domain. In the age of ubiquitous internet and evolving nature of ways people attack minorities, it is important to develop newer methods to assist vulnerable populations. Given the several ongoing crises in the 21st century, tasks similar to the one discussed in this paper are going to be increasingly more important.

Annotating hate speech is a non-trivial task since it is highly subjective [9]. The highlighted paper poses a similar dilemma where a small subset of comments exhibiting *hate speech* towards the (alleged) oppressors were marked as *voice-for-the-voiceless* by the classifier. In such scenarios, defining the scope of the classification task and communicating this information to annotators are crucial. Moreover, these findings underscore the importance of human-in-the-loop both during the Active Learning steps and the subsequent moderation task since complete reliance on automated methods might qualify comments as *voice-for-the-voiceless* that are offensive to other communities.

REFERENCES

[1] Josh Attenberg, Prem Melville, and Foster Provost, 'A unified approach to active dual supervision for labeling features and examples', in *ECML/PKDD*, pp. 40–55. Springer, (2010).
[2] Mladen Dimovski, Claudiu Musat, Vladimir Ilievski, Andreea Hossmann, and Michael Baeriswyl, 'Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings', *arXiv preprint arXiv:1802.00757*, (2018).
[3] Katie Hunt. Rohingya crisis: How we got here, 2017. [Online; accessed 12-May-2019].
[4] Karsten Müller and Carlo Schwarz, 'Fanning the flames of hate: Social media and hate crime', *Available at SSRN 3082972*, (2018).
[5] John T. Nockleby, 'Hate speech', *Encyclopedia of the American constitution*, **3**(2), 1277–1279, (2000).
[6] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi, 'Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features', in *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, (2018).
[7] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell, 'Hope speech detection: A computational analysis of the voice of peace', *CoRR*, **abs/1909.12940**, (2019).
[8] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell, 'Voice for the voiceless: Active sampling for finding comments supporting the rohingyas', in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, p. To appear, (2020).
[9] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki, 'Measuring the reliability of hate speech annotations: The case of the european refugee crisis', *arXiv preprint arXiv:1701.08118*, (2017).
[10] Anna Schmidt and Michael Wiegand, 'A survey on hate speech detection using natural language processing', in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10, (2017).
[11] Vikas Sindhwani, Prem Melville, and Richard D Lawrence, 'Uncertainty sampling and transductive experimental design for active dual supervision', in *Proceedings of the 26th ICML*, pp. 953–960. ACM, (2009).
[12] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski, 'Exposure to hate speech increases prejudice through desensitization', *Aggressive behavior*, **44**(2), 136–146, (2018).