Don't Patronize Me! Data Statement

Carla Pérez-Almendros, Luis Espinosa-Anke and Steven Schockaert
School of Computer Science and Informatics
Cardiff University, UK
{perezalmendrosc,espinosa-ankel,schockaerts1}@cardiff.ac.uk

Version 1.1.

INTRODUCTION

Today, most NLP projects targeting misinformation and unsafety in online conversations focus on explicit, aggressive and flagrant phenomena. These include fake news detection [4]; trust-worthiness prediction and fact-checking [1, 2]; modeling offensive language, both generic [12], and geared towards specific communities [3]; or rumour propagation [6].

However, there exist subtler but equally harmful behaviours in digital media which, unfortunately, are typically overlooked in NLP research and online safety initiatives. In this project, we propose to bridge this gap by tackling the problem of patronizing and condescending language (PCL). An entity engages in patronizing communication when its use of the language shows a superior or condescending attitude towards others. These attitudes, when normalized, routinize discrimination and make it less visible [9].

Research in sociolinguistics suggests the following traits of PCL: (1) It fuels discriminative behaviour by relying on subtle language; (2) it creates and feeds stereotypes, which then drive to greater exclusion, rumour spreading and misinformation [10]; (3) it strengthens power-knowledge relationships [7]; and (4) it is a sign of pornography of poverty [8], a communication style that explicitly depicts vulnerable situations to move a target audience to action. PCL-based communication can be typically found in newswire, social media and political discourse, but also, and often unintentionally, in NGO campaigning.

A. CURATION RATIONALE

We present Don't Patronize Me!, an annotated dataset with Patronizing and Condescending Language (PCL) towards vulnerable communities. This annotated data is especially aimed at the NLP community in order to help improve the modelling and detection of PCL when referring to vulnerable communities, with the ultimate goal of producing and consuming a more responsible and inclusive communication.

The Don't Patronize Me! dataset (v.1.1) consists of 7,638 paragraphs about vulnerable communities extracted from news stories from the News on Web (NoW) corpus [5]. This original corpus contains more than 18 million articles crawled from online media in 20 English-speaking countries (see Table 1) from 2010 until 2018.

In order to create our own dataset, we automatically selected from the NoW corpus just those articles where at least one word from a list of selected keywords was present (see Table 2). The articles were then divided per country and keyword and split into paragraphs. With the objective of assuring a balanced representation of countries and keywords, we randomly selected 75 paragraphs per keyword and country using the SciKitLearn library [11]. The final dataset will be a collection of 15,000 paragraphs with PCL annotations referring to vulnerable communities (150 per keyword; 750 per country).

Countries represented in the Don't Patronize Me! dataset			
Australia	Hong Kong	Sri Lanka	Pakistan
Bangladesh	Ireland	Malaysia	Singapore
Canada	India	Nigeria	Tanzania
United Kingdom	Jamaica	New Zealand	United States
Ghana	Kenia	Philipines	South Africa

Table 1: Countries represented in the Don't Patronize Me! dataset

The keywords include seven potentially vulnerable groups which are widely referred to in general media and are potential recipients of condescending treatment. The remaining three keywords are concepts usually used to describe the former communities or the situations they live.

Keywords				
Disabled	Homeless	Immigrant	Migrant	Poor families
Women	Hopeless	Vulnerable	In need	Refugee

Table 2: Keywords represented in the Don't Patronize Me! dataset

B. LANGUAGE VARIETIES

The paragraphs included in the Don't Patronize Me! dataset are written in English. Twenty English speaking countries are represented in the dataset, thus all their varieties of English are expected to be present in the corpus. Table 3 shows the codes of the English varieties as recommended in BCP-47[1].

It is not possible for us to know either the regional varieties of English in each country, if any, or if English is the speaker's first language.

Language varieties in the dataset			
en-AU	en-HK	en-LK	en-PK
en-BD	en-IE	en-MY	en-SG
en-CA	en-IN	en-NG	en-TZ
en-GB	en- JM	en-NZ	en-US
en-GH	en-KE	en-PH	en-ZA

Table 3: Language varieties represented in the dataset

C. SPEAKER DEMOGRAPHIC

As the paragraphs of our dataset are extracted from another corpus, we do not have the possibility to trace socio-demographic data of the speakers. Nevertheless, we can assume a) they are journalists, as they work in the media, so they are educated professionals; b) they speak English, although we do not know if this is their first language, and c) there is a wide representation of different ethnic origins, as we collect texts from 20 countries.

In our dataset, each country contributes with 750 articles, so this is the maximum number of different authors we could have per country. We have not observed any disorder of speech, as the texts are written, probably edited and reviewed before their publication.

1 D. ANNOTATOR DEMOGRAPHIC

The annotators who collaborated in this dataset are three white females, with ages between 25 and 35 years old. Their first language is Spanish, but they are bilingual in English. They all have graduate and postgraduate studies in communication, computer science and data science.

E. SPEECH SITUATION

The news stories from where the paragraphs of our dataset are extracted were published between 2010 and 2018 in 20 countries (see section A). The stories are asynchronous communication, written, edited and probably reviewed before publishing. We assume the texts are likely intended to reach a general audience, although the characteristics of the audience might vary depending on the country and publishing media.

F. TEXT CHARACTERISTICS

The texts of the dataset belong to the journalism genre and the topics have been previously selected to cover the treatment of the media towards potentially vulnerable groups, as explained in section A.

References

[1] P. Atanasova, A. Barron-Cedeno, T. Elsayed, R. Suwaileh, W. Zaghouani, S. Kyuchukov, G. D. S. Martino, and P. Nakov. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv preprint arXiv:1808.05542*, 2018.

[2] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, and G. Da San Martino. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. In *CEUR Workshop Proceedings, Lugano, Switzerland*, 2019.

[1] https://tools.ietf.org/rfc/bcp/bcp47.txt

[3] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, 2019.

[4] N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[5] M. Davies. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. available online at https://www.english-corpora.org/now/, 2013.

[6] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[7] M. Foucault. *Power/knowledge: Selected interviews and other writings, 1972-1977.* Vintage, 1980.

[8] J. Nathanson. The pornography of poverty: Reframing the discourse of international aid's representations of starving children. *Canadian Journal of Communication*, 38(1), 2013.

[9] S. H. Ng. Language-based discrimination: Blatant and subtle forms. *Journal of Language and Social Psychology*, 26(2):106–122, 2007.

[10] D. Nolan and A. Mikami. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70, 2013.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[12] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.