Identifying Contraception Misconceptions Using Sequential Stance Detection on Posts in Filipino Reddit Communities

Daniel Raymond D. Del Rio, Riana Mary Claire G. Lim
Ateneo de Manila University
Quezon City, Philippines
{daniel.delrio, riana.lim}@obf.ateneo.edu

I. CONTEXT

Reddit is a social news website with around 130,000 active communities, or subreddits, specialized in various topics [1]. Reddit is structured such that users, or Redditors, may reply to comments under posts, thus creating subtopics under the original post which are called comment threads [2]. This allows different topics to stem from the original one and encourages diverse discussions among community members.

Filipino Reddit communities frequently engage in discourse on the website and discuss issues such as politics and current affairs, and topics on sex education and health come up as well. Filipino Redditors discuss their frustrations regarding the state of sex education in the Philippines, or the lack thereof, and even try educating other users on sexual health.

Despite the passage of the Responsible Parenthood and Reproductive Health Act in 2012 and the existence of a few reproductive health centers, the Philippines still has a lot of gaps to fill, especially in reproductive health education. The Department of Education (DepEd) has already been urged by the Department of Health to implement comprehensive sex education due to the rising cases of sexually transmitted diseases and teenage pregnancies [3]. However, in the predominantly Catholic context of the Philippines, initiatives on sexual health are difficult to follow through, and the lack of sexual health education from both adults and institutions leaves the Filipino youth dependent on online and potentially unreliable sources.

II. RESEARCH QUESTIONS AND OBJECTIVES

This study aims to explore how the topic of contraception is discussed in an online community of Filipinos, namely the subreddits of r/Philippines and r/SafeSexPH. Posts and comments about contraception will be extracted from these subreddits and classified based on their tone (i.e. whether they are more informative or more emotional). We also want to find out how well-informed Filipino Redditors are about the topic by determining whether what they know about contraception would be considered misconceptions.

Hence, the questions this study aims to answer revolve around these objectives. First, how do Filipino Redditors view contraception? Are they for or against using it? Second, how much of what Filipinos Redditors know about contraception are actually accurate? Third, what tone do they use in their language when having a discourse about contraception? Lastly, is stance detection an accurate method for classifying the user statements as misconceptions or not?

III. SIGNIFICANCE AND SCOPE

Catholicism and the lack of a concrete reproductive health program in Philippine education make discussions on sex and sexual health taboo in the Philippines. The Internet has become the most accessible means for the Filipino youth to learn about sexual health and their sexuality. However, this may be detrimental if inaccurate information is found or circulates online.

Detecting the stance of posts and comments may aid in identifying misconceptions among the statements on contraception. Furthermore, by determining the accuracy (or inaccuracy) of the information that Filipino Redditors know about contraception, we can discover the most common misconceptions among Filipinos, so that future education programs about reproductive health can work towards correcting these.

This study will only cover comment threads in the subreddits of r/Philippines and r/SafeSexPH, as these are spaces where Filipinos are most saturated in the Reddit community and can discuss the state of the country in many aspects. However, comments or posts that are in Tagalog or Taglish will not be analyzed, as no lexicons of Tagalog words are known by us.

IV. METHODS OF DATA COLLECTION

We will be gathering submissions and comments on the subreddits of r/Philippines and r/SafeSexPH using the Python Reddit API Wrapper (PRAW). Using the Python

Pushshift.io API Wrapper (PSAW), they will be filtered according to which posts and comments contain specified keywords. The keywords are terms or phrases related to contraception, such as "condom" or "birth control". Around 1000 most recent submissions and comments containing these keywords will be collected from the aforementioned subreddits. Other metadata that will be stored aside from the text itself include the author, date and time, and the "karma" of the post/comment.

Since the submission and its comments follow a hierarchical, or like a tree, structure, each entity will also be assigned a depth. The root of the tree, the submission, will be assigned a depth of 1. Its parentless comments will be assigned depth 2, while their children will be assigned depth 3, and so on.

V. PRELIMINARY METHODS OF ANALYSIS

The analysis will be built on the work of Kochkina et al. [4] and Zubiaga et al. [5] exploiting the tree structure of Twitter conversations for sequential stance classification.

A. Feature Extraction

The dataset obtained from the subreddits will first be pre-processed. After the data is pre-processed, the following features will be extracted:

1. Punctuation
    a. Binary feature indicating the presence of question marks
    b. Binary feature indicating the presence of exclamation points
2. Formatting
    a. Length of comment or post
    b. Binary feature indicating the presence of external links or sources
3. Statement Role
    a. Binary feature indicating if the statement is the root
    b. Integer feature indicating the comment depth assignment
4. Statement Similarity
    a. Cosine similarity with root statement
    b. Cosine similarity with the entire thread

B. Classifiers

We will use the models proposed by the study by Kochkina et al., which use the Branch-LSTM Model, "a neural network architecture that uses layers of LSTM units." [4] They were able to incorporate this in their focus of using the tree structure of Tweets similar to Reddit, hence we think we can use this as well for this study.

VI. REFERENCES

[1] Reddit, Inc. (n.d.). The conversation starts here. *Homepage - Reddit.* Retrieved from https://www.redditinc.com/

[2] Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the Reddit community. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*. doi:10.1145/2492517.2492646

[3] Macasero, R. (2018). DepEd urged to implement comprehensive sex education amid rising HIV cases. Retrieved from https://www.philstar.com/headlines/2018/12/04/1874152/deped-urged-implement-comprehensive-sex-education-amid-rising-hiv-cases

[4] Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with Branch-LSTM. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).* doi:10.18653/v1/s17-2083

[5] Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2016) Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* Retrieved from https://www.aclweb.org/anthology/C16-1230