Analyzing Discourse on Contraception in Filipino Reddit Communities

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

ABSTRACT

Sexual health education has been a significant topic in the Philippines, considering the lack of concrete integration of sexual health in the country's education as well as the taboo nature of the topic in this predominantly Catholic country. Some of the Filipino youth have resorted to online communities such as Reddit for learning about sex. This study aims to delve into the discourse about contraception on Reddit and to understand the sentiment of Filipinos on the topic. In creating the dataset, PRAW was utilized for gathering submissions whose titles and self-texts include certain keywords related to contraception. To analyze the comment and submission texts, TFIDF and N-gram vectorizers, Wordcloud, VADER sentiment analysis, and LDA were used. A neural network using an LSTM model was used to classify the stances of the comments based on whether they support, deny, question, or comment on their parent comments to further understand the discourse. We conclude that many Filipino Redditors support the use of contraception based on their stance, sentiment and general word usage and topic discussion in the online community. Vectorizing the parent and child comments as features might not also be sufficient for classifying stance, based on the LSTM model.

KEYWORDS

sexual health education, reddit, stance detection, LSTM

ACM Reference Format:
Daniel Raymond D. Del Rio and Riana Mary Claire G. Lim. . Analyzing Discourse on Contraception in Filipino Reddit Communities. In *CS 193.1: Introduction to Social Computing*. ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

1.1 Context

Reddit is a social news website with around 130,000 active communities, or subreddits, specialized in various topics (Reddit, Inc.). 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 (Weninger, 2013). 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 (Macasero, 2018). 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.

1.2 Research Questions and Objectives

This study aims to explore how the topic of contraception is discussed in an online community of Filipinos, namely r/Philippines. Posts and comments about contraception were extracted from this subreddit, analyzed, and classified based on their stance (i.e. whether they support, deny, question, or comment on the previous comments).

This study aims to understand how Filipino Redditors view contraception and whether or not they are for or against its use. It also takes a look at what words they most commonly use when talking about contraception online.

1.3 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.

Understanding the sentiment and discourse that takes place online would help us understand what the needs of Filipinos are in terms of sexual education and sexual health. For instance, the frequent questions that Redditors ask about birth control pills or intrauterine devices (IUDs) can be relayed to the proper departments and educational units, so they know which topics they should target and include in future curricula about sexual health education.

This study only covered comment threads in the subreddit r/Philippines as it is a space 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 contain Tagalog words may have

CS 193.1, July 22, 2019, ADMU

different or even inaccurate analysis due to the absence of Tagalog words in the corpora we use.

2 LITERATURE REVIEW

Most studies on stance detection on Twitter classify tweets as single units. Lukasik et al. (2016) moved away from this, and analyzed tweets considering temporal or time-related information to put the tweets in context. He added a fourth label which indicates a tweet that did not add anything to the discussion and simply comments on the rumors. Their approach used Hawkes Process (HP) for the sequence classification of stances, and their results supported the idea that using both temporal and textual information will improve model performance (Lukasik et al., 2016).

Zubiaga et al. (2016) exploited the nested tree-structure of Twitter conversations to classify tweet stances. They also used the four-way classification and classified tweets as either supporting, denying, questioning, or commenting (SDQC). They used Conditional Random Fields (CRF) which allowed them to model Twitter conversations as graphs, and their approach showed that considering the discursive interactions on Twitter can lead to significant improvements in classifying the stances (Zubiaga et al., 2016).

Building on the approach to use the hierarchical conversation structure of Twitter, Kochkina et al. (2017) proposed Branch Long Short-Term Memory (Branch-LSTM) which uses "layers of LSTM units" and allowed them to incorporate the entire conversation context.

The work of Tai et al. (2015) includes a more detailed explanation of their tree-structured LSTM model. Their work generalizes LSTM, which is usually used in linear chains, to tree-structured data (Tai, 2015). They proposed two extensions to the sequential LSTM architecture which are the Child-Sum Tree-LSTM and the N-ary Tree-LSTM (Tai, 2015). However, their paper uses the tree-LSTM on sentiment classification of sentences and sentence pairs only, and not on networks of sentences such as those in Twitter.

Since most research work on stance detection has been done on Twitter, we want to apply the tree-structure stance detection on Reddit. While Twitter has a 280 character limit, Reddit has none, which allows Redditors to provide longer comments and replies. Moreover, Reddit also uses a similar conversation structure as Twitter.

We also want to apply the stance detection regarding contraception and contraception use instead of rumors, although the two would require a more or less similar approach for classification and the SDQC classification would still apply.

Moreover, there is a scarcity of social computing papers that analyze discourse on contraception, which makes this an even more significant and relevant topic, especially considering the social context of Filipinos.

3 METHODOLOGY

3.1 Dataset

The Reddit API makes use of a specific syntax for searching posts, considering their fields and attributes:

```
[field]:[query] [AND/OR] [field]:[query]
```

Only the title and selftext fields were used, where selftext is the text in the body of a post, if the submission is a self-post. Keywords related to contraception were first identified to narrow down our queries when searching through the two subreddits ("Glossary of Contraception Term"). In order to make the dataset, the Python Reddit API Wrapper (PRAW) was utilized to gather submissions whose titles or body texts contain the specified keywords.

Certain attributes of all the comments of the submissions were then included in a Pandas DataFrame, namely the comments' author, body text, depth (level at the tree structure), parent_id (ID of the parent comment, if present), id, comment karma, and subreddit. The DataFrame was then converted into a CSV file to be used later on.

Data was labeled manually as either supporting the main comment, denying the main comment, questioning the main comment, or simply commenting on the discussion.

Before the comment texts and the submission texts were fed into the model, they first underwent some preprocessing. Specifically, punctuation and other special characters in these texts, as well as stopwords included in the NLTK corpus, were removed. After that, all the words were lemmatized using TextBlob Lemmatization.

3.2 Text Analysis

The comment and submission texts were represented as vectors using N-gram and Term Frequency Inverse Document Frequency (TFIDF). By extracting the frequencies from these vectorizers, we wanted to identify which words and two-word phrases were most commonly used in all the texts. The results were visualized using Seaborn's bar plot. We also visualized the most commonly used words using a WordCloud.

Aside from frequency, the sentiments of the texts were also analyzed using VADER (Valence Aware Dictionary and sEntiment Reasoner). We preferred VADER over the polarity and subjectivity analyzer of TextBlob because VADER specifically caters to social media texts, thus it takes into account the texts' punctuation, capitalization, and usage of intensifiers, i.e. magnitude of intensity would increase by using exclamation marks, uppercase words, and words such as "extremely" (Randey, 2018). The sentiments of the comments were plotted according to their labels using a histogram and Seaborn's violin plot. We also visualized the relationship between the sentiments of the comment and their parent.

Lastly, the main topics of the texts were also extracted using Latent Dirichlet Allocation (LDA) to determine the most frequently discussed topics and which words were written with regards to those topics. This was then visualized using the pyLDAvis package.

3.3 Model

The labels were converted from strings into integers and then to categorical data in the form of a 4 by 1155 (the shape of the final, cleaned dataset) binary class matrix. The comments and parent data were vectorized using Doc2Vec from the Gensim library. These two vectors were appended together, converted into a Numpy array, and split into train and test data. The training data were then reshaped to fit the LSTM model.

Considering the skewness of the dataset, label weights were adjusted accordingly to prevent the model from overfitting to the commenting label. Comments were given a weight of 12.2, supporting texts were given a weight of 22.9, denying texts were given a weight of 40.1, and questioning texts were given a weight of 50.3.

The model was trained on 100 epochs with a batch size of 64. After training, it was evaluated against the test dataset.

4 RESULTS AND DISCUSSION

4.1 Dataset

From the 200 threads gathered whose titles and self-texts contained the keywords, a total of 3,341 comments were extracted. In order to minimize manual labeling, rows with comments that have a length of less than 200 along with the children of the comments (i.e. all replies of the comments) were removed. There were also submissions that were not related to contraception at all, but contained the specified keywords such as pills. The comments of these submissions were removed. This resulted to a total number of 1,155 comments left.

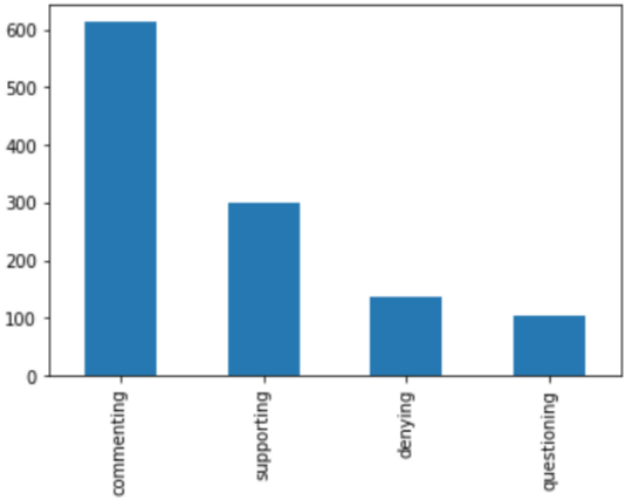

Figure 1: Distribution of labels among all comments.

As expected, majority of the comments were only commenting on their source comment. The next most frequent label was the "supporting" label, while we discovered that the comments who were questioning their parents were the fewest. Some comments included questions about the topic on hand, but we labeled some of them as "commenting" instead of "questioning" because they were merely probing for more information rather than actually questioning the rationale of the parent comment.

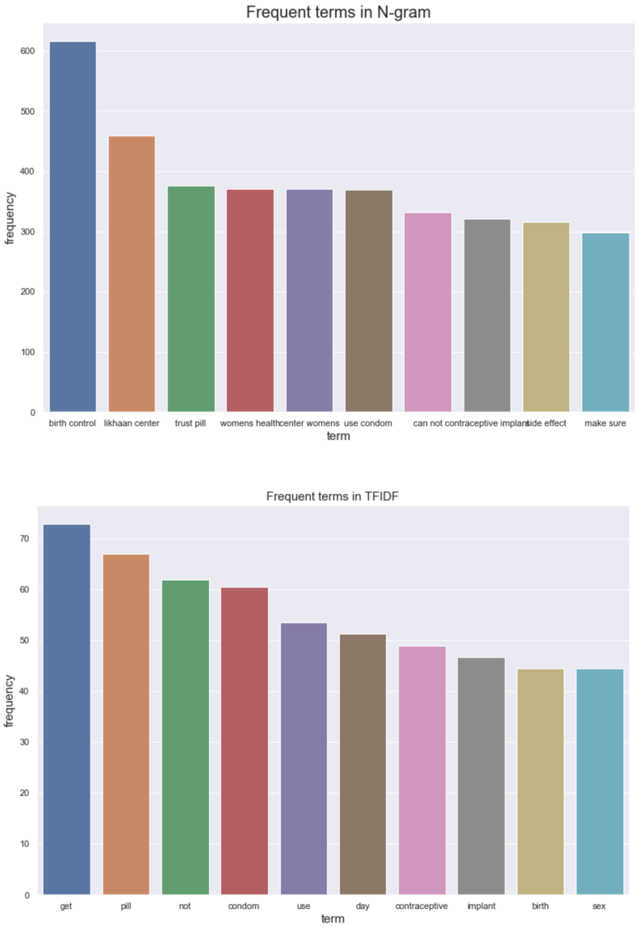

Figure 2: Frequency of terms in submissions using TFIDF and N-gram.

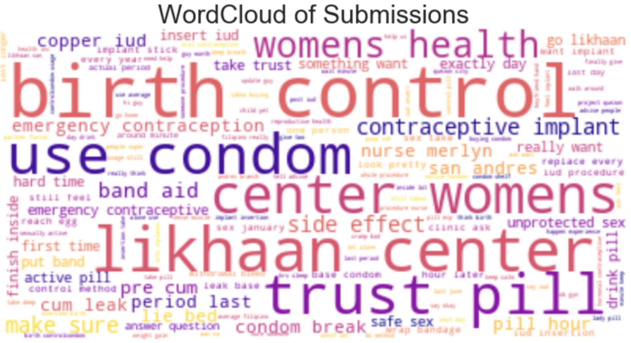

Figure 3: Word cloud of submission texts.

4.2 Text Analysis of Submissions

Figures 2 and 3 show the most frequently used words and phrases in the texts of the submission. Many of the posts either narrate their stories with contraception, such as their IUD insertions in Likhaan Center for Women's Health (hence why sights of Nurse Merlyn and Ate Chona, nurses in the center, appear in the word cloud), or inquire about the selling of contraceptives in convenience stores and health centers such as birth control pills (or trust pills), condoms, and emergency contraceptive pills. Others also ask for advice about what they should do after their sexual experiences for them or their partners to avoid getting pregnant, which is why "use condom" came up a lot.

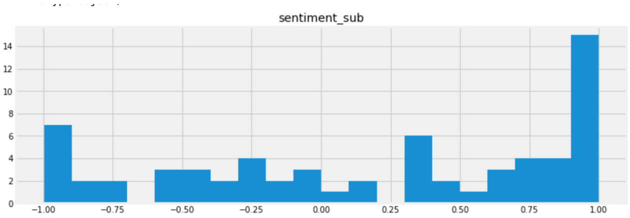

Figure 4: Histogram showing the frequencies of sentiments of the submission texts.

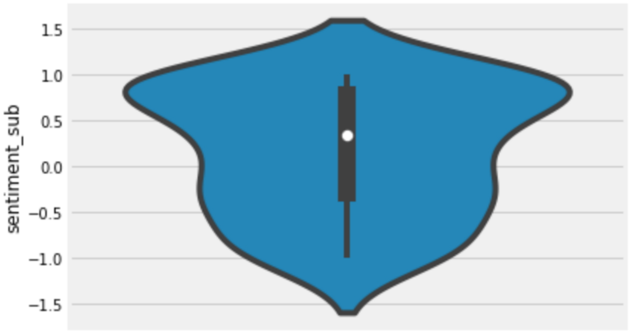

Figure 5: Violin plots showing the distributions of sentiments of the comment texts per label.

We can see here that many of the submissions have positive sentiments around the 0.75 - 1.00 mark, with the next most frequent sentiment in the neutral range. The high sentiment can be attributed to the typically encouraging and grateful tone reflected in the language they use in their posts. For instance, many posts with the highest sentiments end their posts with a note thanking Reddit for their advice and insights regarding the issue. They also narrate their stories in the health centers with positivity, using words like "extremely happy" and encouraging others to do the same.

Meanwhile, the posts with negative sentiments narrate their unpleasant experiences at convenience stores (i.e. being ID'ed when buying condoms) or criticizing the tabooness of contraception in the country. Others also post in a panic state as they ask for advice about what to do with their partners after having unprotected sex, so they tend to use sad faces.

For the topic modelling based on the pyLDAvis of the N-gram and TFIDF vectors, we inferred that the topics are related to their experiences at Likhaan Center, experience buying contraceptives at convenience stores, narrations of their sexual experiences, and their experience with contraceptives just after their sexual experience. These topics are found in the visualization of both vectors, but their circles differ in sizes. However, in the TFIDF vector, there is a small circle containing many Filipino words, which were probably grouped together because the model was unable to properly characterize them. Nevertheless, the cluster was relatively small, indicating that not many Tagalog words were used, something differing from the cluster in the pyLDAvis of the comments. There were also a few Tagalog phrases in the N-gram visualization, implying that when Tagalog words were used, they appeared in tandem with English words (i.e. Taglish).

4.3 Text Analysis of Comments

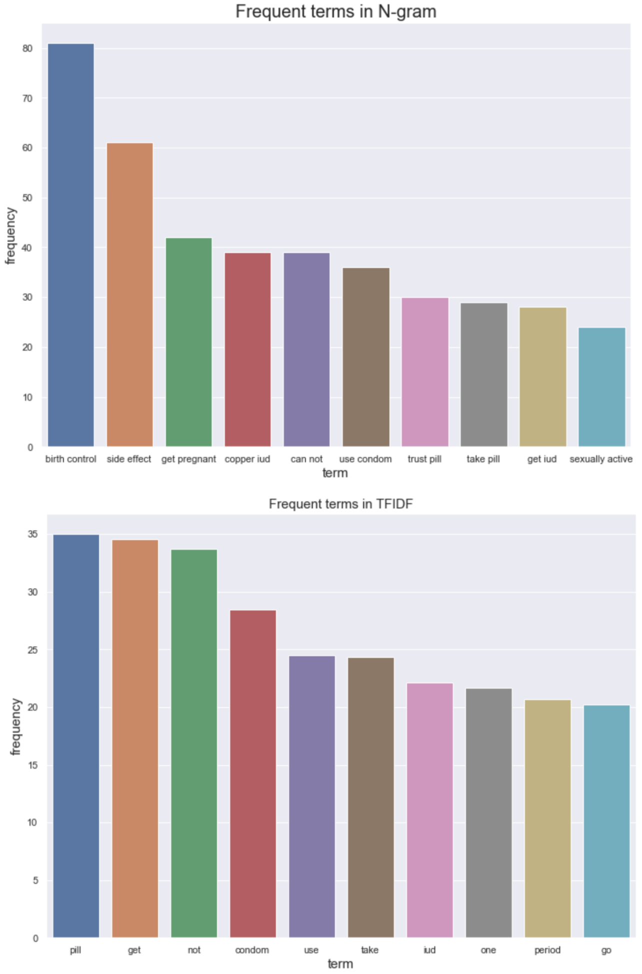

Figure 6: Frequency of terms in comments using TFIDF and N-gram.

Figures 6 and 7 show the most frequently used words and phrases in the comment text. Many of the comments talk about the use of contraception itself. Of the different contraception options, "pill", "condom", "iud", and "implant" show up the most. We believe that the frequency of these contraceptives could be representative of their availability. Condoms and pills are much easier to access for Filipinos because they can be purchased from stores. Although condoms are much more convenient to use because pills requires consistent daily use, ordinary contraceptive pills also serve another purpose as emergency contraception, or the "plan B" pill, which is now unavailable to obtain legally in the Philippines. Many Redditors mentioned the "Yuzpe" method which is a form of emergency

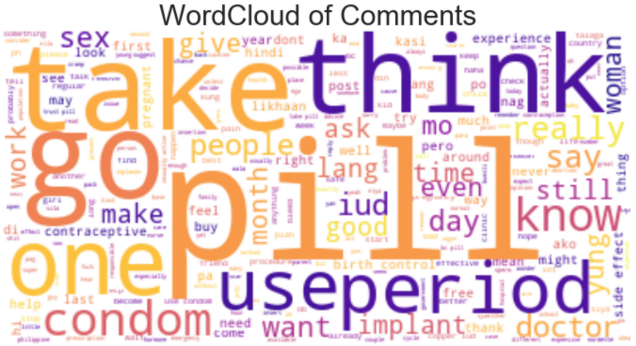

Figure 7: Word cloud of comment texts.

contraception where a woman takes 4 pills within 72 hours of intercourse, and 4 more pills after 12 hours. This may also be why the word "takes" comes up frequently. As for IUDs (intrauterine devices) and implants, these are available at the Likhaan Center which was frequently mentioned in the submission text. These two are much more convenient to use because they are long-term contraceptives, unlike pills and condoms which have to be used consistently. However, they can be difficult to obtain especially for younger women who might be apprehensive about going to the health center.

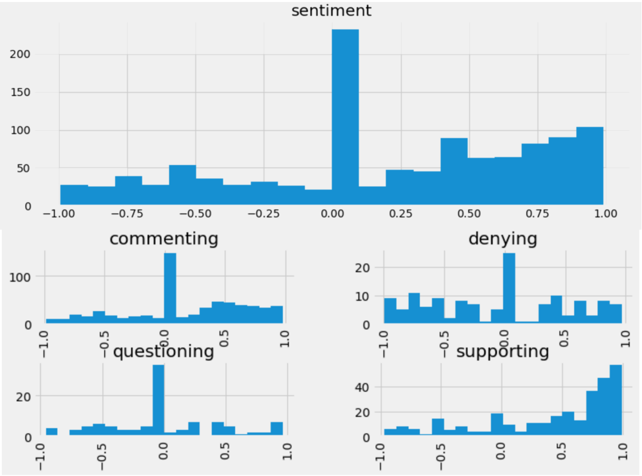

Figure 8: Histograms showing the frequencies of sentiments of the comment texts.

Figures 8 and 9 show how the sentiment of the comment texts are distributed. The sentiment is generally positive, which indicates that the comments are more supportive environment in the comments sections of the posts. A lot of commenters praise the creator of the posts for practicing safe sex and using contraceptives. Comments that support their parents have a much more positive sentiment, whereas comments that deny their parents have a more negative sentiment.

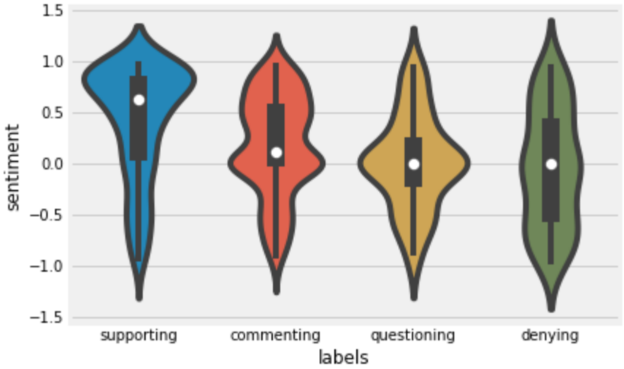

Figure 9: Violin plots showing the distributions of sentiments of the comment texts per label.

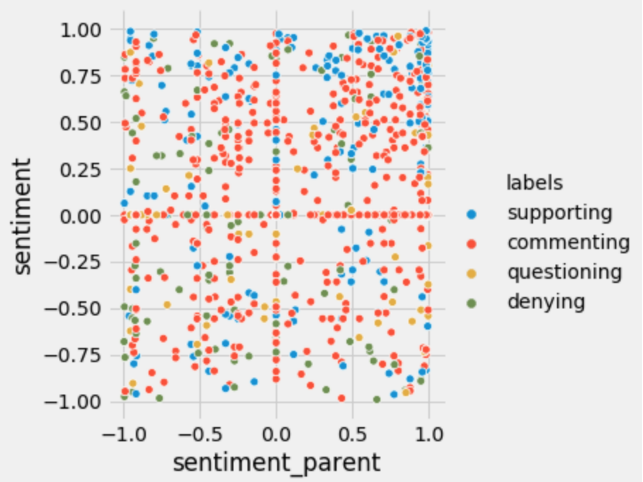

Figure 10: Relational plot showing the sentiments of the comment and its parent.

Figure 10 shows the relationships between the sentiments of comments and their parents. For comments labelled commenting, questioning, and denying, there is generally no relationship. However, for comments labelled supporting, some clustering can be observed where the sentiment of both are positive. We can say that supportive parent comments generate supportive child comments.

For the topic modelling based on the pyLDAvis of the N-gram and TFIDF vectors, we inferred that the topics are related to support and thanking post-writers for sharing their experiences, side effects of different birth control methods, experiences with the contraceptives themselves, and pregnancy scares. The topic related to thanking writers indicate more support for the use of contraceptives. In that topic we saw phrases such as "thanks share", "young people", and "share experience" which might be encouraging more sex education awareness, especially for the youth. For the topic on birth control side effects, we saw phrases such as "side effect", "not sure", "mood swing", and "irregular period". These could indicate the common concerns and difficulties of Filipinas with their different birth control options. IUDs and pills are also frequently mentioned in this topic, which might indicate that these cause the most side effects for them. We also observed mentions of going to the OB-GYN, which might indicate that some concerns should be brought up, or that they should consult with medical professionals regarding the best options for them. The topic regarding their experiences with contraceptives generally just enumerate the different contraceptive options such as condoms, pills, the calendar method, and implants. Emergency contraception and pregnancy scares was the third largest topic in the LDA, which might be because people don't know where else to go for advice regarding this matter. There were mentions of "pill ecp" (emergency contraceptive pill), "unprotected sex", and "copper iud". As mentioned previously, pills can also function as emergency contraception. However, the Yuzpe method can cause mood swings and heavy periods which is also mentioned in the topic. IUDs can also function as emergency contraception if inserted a few days after unprotected sex. There was also a topic which with a more sarcastic tone about congratulating post-writers and commenters on the possibility of becoming a dad.

4.4 Model

The LSTM classifier was tested against both the test data and the entire dataset 10 times. On average, the model performed with an accuracy of 0.41 and a loss of 1.31 on the test data, and with an accuracy of 0.44 and a loss of 1.30 on the entire dataset. At best, the model performed with an accuracy of 0.49 and a loss of 1.26 on the entire dataset.

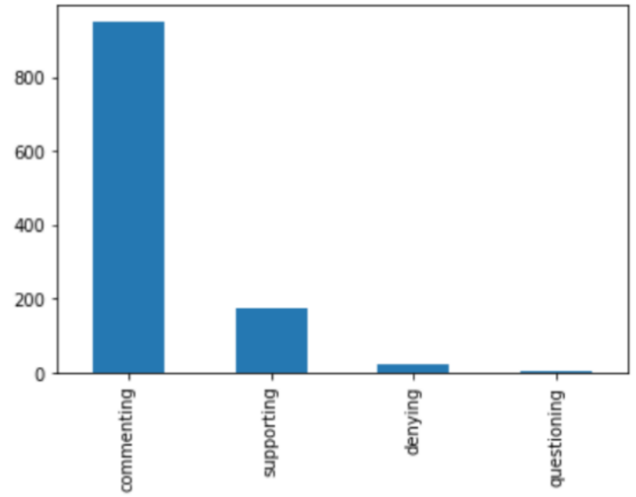

Figure 11: Bar graph showing the distribution of the labels classified by the LSTM model.

However, the best iteration on the test data had an accuracy of 0.47 and a loss of 1.33, but performed a bit poorly when tested against the entire dataset because it overfit to prefer the commenting or supporting labels.

Overall, the model performed poorly with the vectorized parent and child comments as features.

5 CONCLUSIONS AND RECOMMENDATIONS

We conclude that many Filipino Redditors support the use of contraception, evidenced by the ubiquity of posts and comments discussing, inquiring about, and using contraception in the subreddit r/Philippines. If not neutral, many of the sentiments of the comments and posts gear towards the positive side, due to the Redditors' encouraging and grateful tone when posting and replying about contraception. With regards to their stances, majority of the comments seem to be merely commenting on their parent comments, but many comments also seem supportive of their parent comments.

The model also indicates that the vectorized parent and child comments as features might not be sufficient for classifying the stance. Possibly, a recursive model might perform better considering the hierarchical structure of Reddit comments. Identifying stance would not only help us understand Reddit discourse on contraceptives, but may also be applied to detecting misconceptions on the topic.

After this study, we recommend that researchers make use of word vectors and corpora containing Tagalog words, in order to further improve the topic modelling, sentiment analysis, and prediction using LSTM. An example of such is the fastText Tagalog models, which can be used to represent and even classify texts. Furthermore, a more expansive dataset that includes posts and comments in subreddits specifically for discussions on sexual health, such as r/SafeSexPH, is recommended, so that possibly different demographics can be taken into account and covered by our dataset.

For the model, we also recommend that models other than the LSTM, such as convolutional neural networks (CNNs), are used for comparison of accuracy and loss. For further testing, vectorizers other than Doc2Vec may also be used to see which would yield better representations and thus higher accuracies.

REFERENCES

[1] Glossary of Contraception Terms. (n.d.). Retrieved from https://www.mycontraception.ie/en/SSL/glossary-of-contraception-terms.php

[2] 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

[3] Lukasik, M., Srijith, P. K., Vu, D., Bontcheva, K., Zubiaga, A., Cohn, T. (2016). Hawkes processes for continuous time sequence classification: An application to rumour stance classification in Twitter. Proceedings of the 54th Meeting of the Association for Computational Linguistics. doi:10.18653/v1/P16-2064

[4] 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

[5] Randey, P. (2018, Sept 23). Simplifying Sentiment Analysis using VADER in Python (on Social Media Text). Retrieved from https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

[6] stanfordnlp. (2015). treelstm. Retrieved from github.com/stanfordnlp/treelstm

[7] Tai, K., Socher, R. Manning, D. (2015). Improved Semantic Representations FromTree-Structured Long Short-Term Memory Networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. doi:10.3115/v1/P15-1150

[8] 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

[9] 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

6 APPENDIX

You may find our code here: https://github.com/daniddelrio/social-computing