GENDER AND EXTREMISM ON TWITTER IN THE ASIA AND PACIFIC REGION *[PILOT STUDY]*

In this study, we analyze tweets from Bangladesh for the past 6 months. We extract keywords and explore their frequency. Then we move on to our subject matter and focus on tweets related to woman/women. We explore the keywords associated with the words woman/women. Finally, we produce a visualization of the keywords and their frequencies in the form of word clouds. The output of this analysis is as follows:

- Top 100 keywords from the tweets.
- Top 100 keywords associated with the word 'woman'.
- Top 100 keywords associated with the word 'women'.
- Top 500 keywords associated with both ‘woman’ and ‘women’.
- Word cloud visualization of all tweets.
- Word cloud visualization of tweets that contain the words 'woman'/'women'.
- A visual comparison of the proportion of tweets that are focused on the subject matter.

Methodology

The analysis was carried out in python language(version 3.5) in a jupyter notebook[1]. The NLTK library was used for text processing which is the de-facto natural language processing library in python[2]. It has a dedicated library for processing tweets. [3]

1. In this study, we analyzed only the tweets that were posted in English language. So, at first, we extracted the English tweets from the dataset.
2. Each tweet was tokenized and frequently used words like a, an, the, etc. which are commonly known as stopwords were excluded from analysis.
3. We created a twitter specific stopword list and these stopwords were also excluded from analysis.
4. Twitter is a microblog and texts are limited to 140 characters only. So, stripping stopwords leaves us with the words that carry significance, specially adjectives, nouns, adverbs, interjections, prepositions, and verbs excluding conjunctions, pronouns, and articles and emoticons.

[1] http://jupyter.org/
[2] http://www.nltk.org/
[3] http://www.nltk.org/_modules/nltk/tokenize/casual.html#TweetTokenizer

5. Then we counted occurrence of each word or token in the whole tweet corpora with the help of NLTK library and python code.

6. Then we sorted the words in a descending order of frequency and extracted the top 100 words that are most frequent in all the tweets.

7. Our next objective was to extract keywords that are associated to woman/women in the tweets. To achieve this objective, we kept two separate counters that held the keywords that occurred with the words woman and women respectively. We also maintained another counter object that combined the associated keywords for both the words.

8. We scanned through every tweet, searched for the word woman/women in them and when found, we updated the respective counters accordingly with the associated keywords from that tweet. The counter object is specifically designed for this purpose and with every update it aggregates the count of the keywords generated from the tokenization and stopword stripping process.

9. So, with an end to end scan of the dataset, the counters gathered all the keywords available and their frequencies over the whole dataset.

10. Then we sorted the counters in decreasing word and extracted the top 100 words for each keyword and top 500 words for both the keywords combined. The counter was aggregated over the whole dataset, so the extracted top N keywords in each category has a high level of co-occurrence with respective keywords namely, woman and women.

11. Then we combined the whole tweet text corpora (excluding stopwords) for each category and visualized them with word cloud. Size of the fonts of different words in the cloud is proportional to their frequency in the tweets. The map of Bangladesh is used as the mask image for the word cloud.

12. Lastly, we ran a comparison of all tweets to get an idea about the proportion of the tweets that are directed to our subject of concern.

Results

- Total number of tweets in dataset: 8377563
- Total number of tweets in English in the dataset: 6913412
- Top 100 keywords for each category are attached as csv files in tabular format.

- Associated word cloud visualizations are also provided as image files as well as pasted in this document.

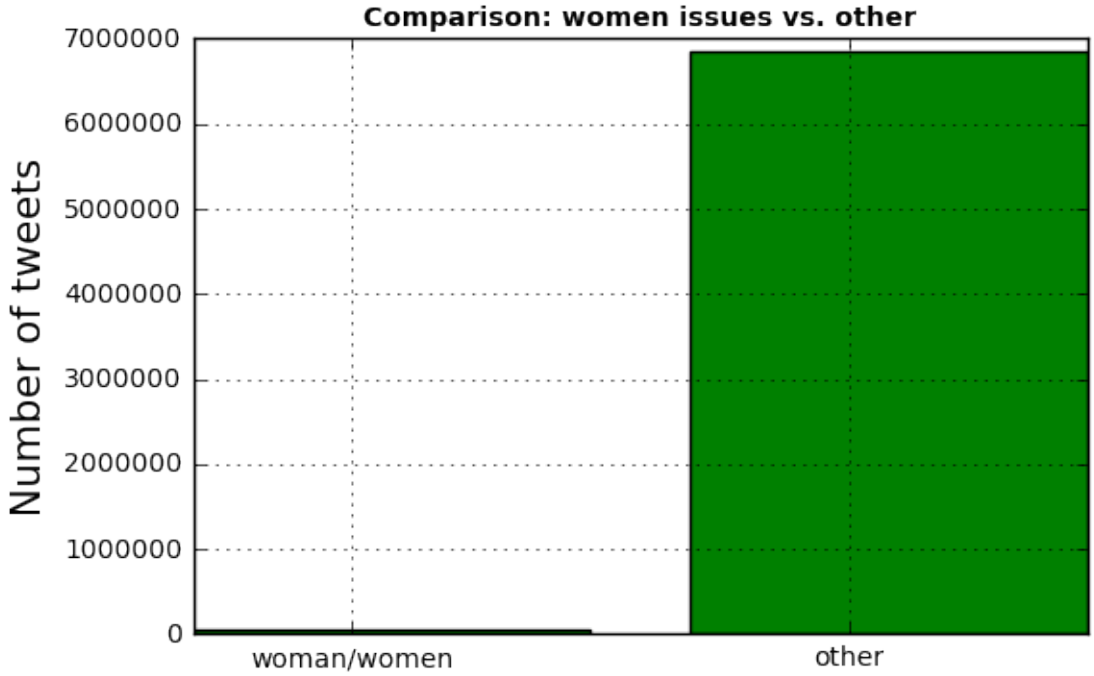
- Visualization in the form of a bar chart of a comparison of tweets directed to the subject of concern to other tweets is also provided as image file. The image is pasted in this document as well for ease of reference.

- The notebook file (.ipynb) is attached with this document.

Word Cloud for tweets containing woman/women

Word Cloud for all tweets from Bangladesh

Number of tweets containing the words 'woman'/'women': 64382
Number of tweets that do not contain these words: 6849030



Comparison of tweets related to subject matter to other tweets

Conclusion and Recommendation

The duration of the collected data coincided with the US election period. That explains the high association with words like Trump and Hillary. Words like men, man, love, beautiful have come up quite frequently. It is important to note that words like attack, rape, violence, nasty, abortion, etc have also come up quite frequently in these tweets which hint to the tensions in gender issues in the region.

Use of twitter is fairly limited in Bangladesh. The portion of Bangladeshis that do use twitter are from a very narrow cross-section of the society, mostly the urban upper, upper-middle class. That explains the fairly modest amount of tweets regarding the subject matter. Focusing on facebook may yield better discovery in this regard as facebook has penetrated nearly all social classes in Bangladesh.