Analyzing Sentiments From Street Harassment Stories

Parvathi Chundi*
Computer Science Department, University of Nebraska-Omaha
6001 Dodge Street, Omaha, NE
pchundi@unomaha.edu

April Corbet†
Computer Science Department, University of Nebraska-Omaha
6001 Dodge Street, Omaha, NE
acorbet@unomaha.edu

ABSTRACT

Street harassment is a pervasive problem that typically targets women and LGBTQ community. There are no effective ways to deal with the harassers as the acts of harassment happen randomly and are difficult, if not impossible, to prosecute. Hollaback! is an international movement aimed at stopping street harassment. Hollaback! servers collect street harassment stories from victims around the globe to share, gather statistics, and create awareness. In this paper, we present a preliminary study focused on analyzing a small sample of Hollaback! stories submitted from New York city. The LIWC software [1] is used to measure the positive and negative emotions hidden in each story and correlate it to the socio-economic status of the location from which the story was submitted.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

algorithms, experiments

Keywords

linguistic inquiry, word count, sentiment analysis

1. INTRODUCTION

Women and LGBTQ individuals are routinely subjected to street harassment simply for being a woman or gay. This gender-based violence happens in every country, with varying degrees of severity. The acts of street harassment are pervasive and do not get reported. Victims bear with it silently since there is no systematic way to reproach the perpetrators as there is for other kinds of sexual harassment. Hollaback! (ihollaback.org) is a movement with a mission to end street harassment around the globe. It is operated by local activists in 50 cities, 17 countries, and 9 different languages. It is a crowd-sourced initiative that empowers women and LGBTQ individuals to break the silence by sharing their stories and pictures. The collected stories are then presented to the local officials where street harassment incidence is high. The movement trains individuals to become leaders in this grassroots movement to end street harassments and conducts educational workshops in local schools and universities to shift public opinion.

There have been many recent works focused on extracting sentiments and opinions from text documents (See [2] for a comprehensive survey). Our long term goal is to use text mining and sentiment analysis to extract actionable information from Hollaback! stories. For example, stories can be separated into categories such as inspirational, victimization, and informational. A story where the actions of a victim resulted in his/her increased confidence can be identified and presented to other users to spread the positive message. Stories of severe victimization can be presented to the authorities for further action. Temporal and spatial characteristics of the stories can be analyzed to identify any trends for cities in the U.S and across the globe.

In this paper, we describe a preliminary analysis of a small sample of Hollaback! stories submitted to the New York city server. We analyzed the sentiments of the stories using the LIWC 2007 [1] software. LIWC analyzes a text document to identify various emotional, cognitive, structural, and process components present in an individual's verbal and written samples to calculate the percentage of words in over 70 language dimensions. It has been used to analyze word usage in poetry of suicidal and nonsuicidal poets [3], emotional content in computer mediated communication, and other applications [5].

We used the LIWC software to measure the amount of words expressing positive, negative, and anger emotions. We also analyzed the demographics of the locations from where the stories were submitted. Based on our analysis of a small data set, we found that the socio-economic status of a person may have no effect on how an individual feels about a harassment incident. We are currently studying the characteristics of stories with high and low positive emotions to see how to distinguish between them.

*Supported by the UNO's Fund for Investing in Research Enterprise (FIRE).

†Supported by the UNO's Fund for Undergraduate Scholarly Experience (FUSE).

2. PRELIMINARY EXPERIMENTS

We obtained a sample of about 2100 stories from the Hollaback servers from which we collected the stories submitted to the New York City server (http://nyc.ihollaback.org). There were about 110 stories in this data set which were analyzed in our study.

Each story in the data set included an ID, time stamp of the submission, a category value (one of the following four values – groping, verbal, stalking, and other), title, edited description of the street harassment incident, latitude and longitude of the location from which the story was submitted.

We analyzed each story with the LIWC 2007 software. For each story, the LIWC 2007 software generates a vector of numeric values for about 70 dimensions. Dimensions include pronouns, personal pronouns, verbs, auxiliary verbs, positive emotions, negative emotions, anger, sadness, among other attributes. For instance, numeric values for positive and negative emotions are non-negative real values, typically in the range 0 to 10.

In our first study, we analyzed the location stamps of the stories and related them to the emotional content. To do this, we first extracted the latitude and longitude values from each story and obtained the approximate postal code for that location using the *geoPlugin* service (www.geoplugin.com). We then used the postal code to obtain demographic information for each story. For this purpose, we used the demographic information recorded on Trulia$_{TM}$ (www.trulia.com). Using the demographic information, we separated the stories into lower and higher socio-economic status. We assigned the ranking based on the median house price of that zip code. If the median house price of the postal code is higher than the median house price of the entire state, then that postal code (and the corresponding story) is assigned a higher rank for socio-economic status. Otherwise, it is assigned the lower status.

Based on these ranking, we found that approximately 23% of the stories were submitted from postal codes of higher socio-economic status. The rest came from postal codes with a lower socio-economic status. This finding is note-worthy in two aspects. It validates the assumption that street harassment affects more people from lower socio-economic status. However, it also underscores the fact the it affects people of all backgrounds.

We then studied the emotional contents of the stories. According to LIWC Online, the measure of negative emotion in personal texts is 2.6 and positive emotion is 2.7. If a story has an LIWC value of more than 4 for positive emotion (or negative emotion) column, then it is deemed highly positive (or negative). On the other hand, a measure of 2 or lower is considered to be low on positive (or negative) emotion.

When we analyzed the LIWC values for stories from lower socio-economic status, we found that almost 80% of them were low on positive emotion. Surprisingly, the negative emotion measure values for these stories were either low or almost equal to that in personal stories (which could be considered as zero negative emotion). Upon careful examination of these stories, we found that the text of a story described the harassment incident with a few sentences expressing anger or disgust, which might have led to a low negative emotion measurement. The rest 20% of the those stories classified as lower socio-economic status, had a high measure for positive emotion. We repeated the above analysis for stories from higher socio-economic ranking. Only about 10% of the stories measured high on positive emotion.

Although we expect most stories to have a low measure on positive emotion, this preliminary study indicates that the socio-economic status of an individual may have no bearing on the how she/he might feel about the harassment incident. We plan to extend the experiments to larger data sets to see if our observations can be validated.

3. CONCLUSION AND FUTURE WORK

Street harassment tends to affect women and LGBTQ individuals disproportionately. Hollaback! is a movement across the globe to curb street harassment, collect statistics, and educate the officials and the public. Our longer term goal is to use text mining and sentiment analysis techniques to analyze the stories submitted to Hollaback servers and to obtain temporal and spatial trends from the data. In this paper, we present a simple study aimed at understanding the relation between the socio-economic status and emotional content of the stories. Our next study will focus on developing methods for identifying positive and inspirational stories that can be prominently featured on the web site.

4. REFERENCES

[1] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic Inquiry and Word Count: LIWC 2007", (www.liwc.net).

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Now Publishers Inc., 2008.

[3] S. W. Stirman, and J. Pennebaker, "Word Use in the Poetry of Suicidal and Nonsuicidal Poets", *Psychosomatic Medicine*, 2001.

[4] J. T. Hancock, K. Gee, K. Ciaccio, and J. M. Lin, "I'm sad you're sad: emotional contagion in CMC", *Proceedings of 2008 ACM conference on Computer supported cooperative work (CSCW2008)*, 2008.

[5] J. W. Pennebaker, "The Secret Life of Pronouns: What Our Words Say About Us", Bloomsbury Press, 2011.