Study of Analytical Tools in Identifying and Preventing Cyberbullying on Anonymous Social Media Platforms

Uffaz Nathaniel
Johns Hopkins University
Baltimore, MD 21218
Email: unathan1@jhu.edu

ABSTRACT

While social media has enabled greater connectivity and communication among its users, it has also been a platform that bullies use to target their victims. Recent studies have shown cyberbullying to be a serious problem among the adolescent. Social media companies therefore have an inherent incentive and responsibility to ensure their platforms are free from cyberbullying, harassment, and objectionable content that includes sex, drugs, racism, gender discrimination, and so on. This paper thus investigates analytical tools such as topic modeling, sentiment analysis, and automatic classification of text to actively monitor anonymous social media platforms, for the purpose of monitoring aforementioned activities. We find cyberbullying and themes of sex and partying to be widespread in the anonymous social media platforms Yik-Yak.

1. INTRODUCTION

Social media plays a central role in modern communication and relationships. The current populous, ranging from teens to the elderly, has either some social media presence or has at one point leveraged a social media service [like Facebook or WhatsApp] to communicate with someone. While adoption of social media services spans the age gamut, statistics show that adolescent and young adults in particular are more likely to use social media, and in more diverse ways (e.g. smartphones, tablets, fitness trackers, laptops) than the elderly [1]. This higher density of adolescent and young adults, coupled with behavioral age dynamics, thus translates into the fact that younger users are more likely to be victims of online harassment and bullying, compared to their older counterparts.

Harassment and bullying itself is not a new phenomenon, however, cyberbullying has evolved as new platforms and technologies have emerged. It should be stated that most social media platforms are not inherently ill. Throughout history any evolution of technology has brought both advances and ills for society. Social media is no exception. It has enabled users to establish new relationships, while also maintaining existing friendships. On the flip-side, when the content is not patrolled, social media has increased the risk of children being exposed to threating situations that include sex, drugs, depression and suicidal posts, discrimination, and cyberbullying [1]. This risk, coupled with the ability to stay anonymous, and reach an audience 24/7, means social media has now become a convenient way for bullies to target their victims at any time.

One type of social media where this cyberbullying and objectionable content is quite pervasive are anonymous question-asking social media sites. These sites include platforms such as Yik-Yak, Whisper, Secret, Nearby, Islands, Sarahah, and YOLO [2].

Despite being platforms with dangerous content, teens and young adults flock to them for anonymous gossip [2]. The diversity and emergence of such platforms is a testament to how captivated teens and young adults are with anonymous platforms.

Yik-Yak, which is the focus of investigation for this paper, is a now defunct platform that allowed people to create and view discussion threads (termed "yaks") within a 5-mile radius [3]. During its prime [at an evaluation of $400M], one of the biggest criticisms of this platform was the widespread cyberbullying and harassment committed through the app. Yik-Yak, unable to curb and address this concern themselves, was the subject of a wide spread ban implemented by various schools and school districts [4]. The inability to contain cyberbullying coupled with inability to address law-enforcement requests ultimately led to the shutdown of the app [3].

The Yik-Yak story shows, it is in the inherit interest of the social media companies to curb cyberbullying and objectionable content in order to prevent a shutdown.

In this paper, we show how topic modeling along with content classification can be used as one of the tools to actively patrol social media communities.

2. RELATED RESEARCH

Given the large volume of data that needs to be analyzed, methods are needed to automatically find themes or "topics" in the content people are talking about. This can enable a moderator to step-in and look at particular communities or threads.

One of the most popular approaches for finding latent (or hidden) topics is the Latent Dirichlet Allocation (LDA) [6]. LDA builds a topic per document model and words per topic model, which are modeled as Dirichlet distributions. In essence, LDA represents documents as "mixtures of topics" that build a document with certain probabilities [8].

However, one of the most common criticisms of LDA is the issue of short text modeling which includes modeling tweets or yaks. Hong et al. in an empirical study demonstrated that topic modeling approaches can be very useful for short text either as solely used features or as complementary features for multiple real-world tasks [5].

Finally, detecting cyberbullying or objectionable content can be approached as a binary classification task where positive instances can represent bullying/objectionable content while negative represent safe posts [1]. We can daisy-chain multiple classifiers one for each objectionable content that includes sex, drugs, depression, suicide, and so on. A positive instance in any one of these classifiers means post should be flagged for a moderator. Building classifiers for each objectionable content continues to be a heavily researched area, and various machine-learning based approaches (both supervised and unsupervised) continued to be applied with an increasing number of success [1].

3. EXPERIMENT

The goal of the experiment performed in this paper was to determine how users of Yik-Yak are using the app, whether they are engaging in cyberbullying or posting questionable content, and ultimately if that can provide us some insight into how users behave on other anonymous social media platforms. Each post within the Yik-Yak ecosystem contains geo-temporal data along with a timestamp that identifies date of postage. The motivation for this paper was thus to leverage this metadata to focus on a specific region, and then analyze the yaks to understand user behavior.

The first step in this experiment was to understand general trends. To extract trends over a specific timeframe, Latent Dirichlet Allocation (LDA) topic modeling was used. The topics were then compared to understand the overall discourse.

The second step in this experiment was to use an existing off-the-shelf classifier to classify posts as marked either as safe or bullying posts.

The third step was to use sentiment analysis to understand the overall mood of yaks.

3.1 Data Collection

Since the Yik-Yak API is now unavailable, we had to rely on pre-compiled data. For this we turned to Nicholas Yager's data of yaks the he collected from Geneseo, NY region[1]. We found the data especially good for one main reason: it is localized to Geneseo, NY which enables us to make inferences about how cyberbullying takes place in local communities—specially schools. Additionally, this dataset has 6663 unique messages that were posted between Jan 1, 2015 and March 4, 2015 making it statically significant.

The second dataset that was used was the Sentiment140 dataset[2]. This dataset contains 1.6 million tweets which have been manually labeled as negative or positive. We leveraged this dataset to build a sentiment classifier that enabled us to classify yaks as positive or negative.

Lastly, on off-the-shelf classifier built at the University of Wisconsin-Madison was leveraged to classify whether a post contained traces of bullying.

3.2 Data Pre-Processing

To prepared the data for topic modeling, we needed to 1) clean the Yik-Yak dataset, and 2) tokenize it.

To process the messages, the Natural Language Toolkit (NLTK) in Python was leveraged. Messages were first stripped away of any HTML entities, punctuation, and any special characters that do not include emojis. Each message was then tokenized using the ToktokTokenizer tokenizer. The Tok-Tok tokenizer is ideal for our dataset because it is a simple, general tokenizer, where the input has one sentence per line. It is also extremely fast. After tokenization, NLTK's English stop words along with the other commonly stop words found through exploratory data analysis were removed. Lastly, words that appeared only once in our corpus were removed to help LDA produce better mixtures of topics.

3.3 Data Exploration

We began by comparing our Yik-Yak dataset to Twitter Sentiment140 dataset. We wanted to see if there are any similarities among the micro-blogging platforms. Results are shown in Figure 1.

[1] https://gitlab.com/nicholasyager

[2] http://help.sentiment140.com/for-students

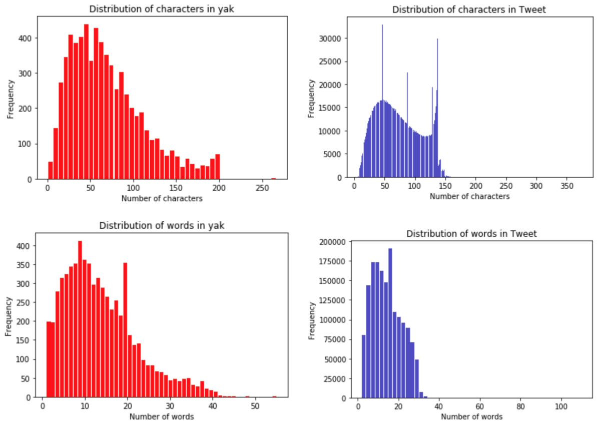


Figure 1. Distribution of character counts and word counts in Yik-Yak (red) and Twitter (Blue). The mean character count for Yik-Yak is 70 and Twitter is 74. The mean word count is 14 for both platforms. This means both platforms are similar and similar analytical tools can be used on Yik-Yak dataset.

The next thing we looked at was to gauge the overall sentiment. To accomplish this, the Sentiment140 dataset was used to build a Logistic Regression classifier. To vectorize our corpus, a Term Frequency-Inverse Document Frequency (TF-IDF) model was used. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Thus the more often a word is used (e.g. stop-words), the lower its weight. Conversely, the more rare a word is, the higher its weight. This classifier achieved a 89% accuracy. Results for sentiment analysis on Yik-Yak dataset are illustrated in Figure 2.

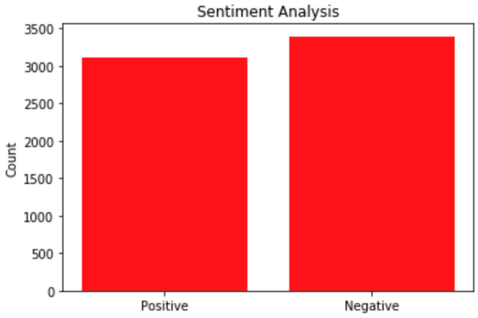


Figure 2. Sentiment analysis on Yik-Yak dataset. Out of a total of 6507 yaks, 3111 were classified as *positive* and 3396 as *negative*.

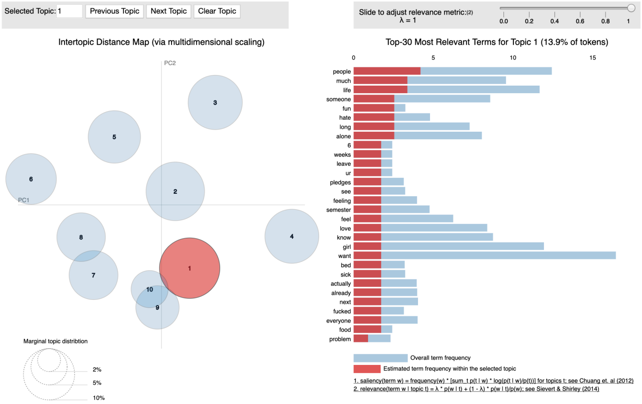
At this point, a LDA model was built using the Gensim[3] toolkit for Python. The model generated topics using both a bag-of-words representation for corpus as well as a TF-IDF weighted corpus. Topics were generated for the whole dataset as well as smaller subsets that were partitioned by week (e.g. Valentine's Day week from Feb 8, 2015 through Fab 15, 2015). The hope was to see any changes in topics over the course of days or weeks.

Figure 3 shows a plot generated for 10 topics using the pyLDAvis[4] visualization tool and Table 1 shows the topics for Valentine's Day week in 2015.

In total ten different topics were generated where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic.

A tool like pyLDAvis is especially useful in assessing the quality of our topics. A good topic model generally has fairly big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant [7].

A model with too many topics, will typically have many overlaps, small sized bubbles clustered in one region of the chart [7]. We see that in Figure 3, while there was some overlap, the topics were fairly large, scattered through the chart – indication thus a sufficiently good topic model.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3 – LDA Visualization of topics in the week of Feb 8, 2015 through Fab 15, 2015 using TD-IDF weighted corpus.

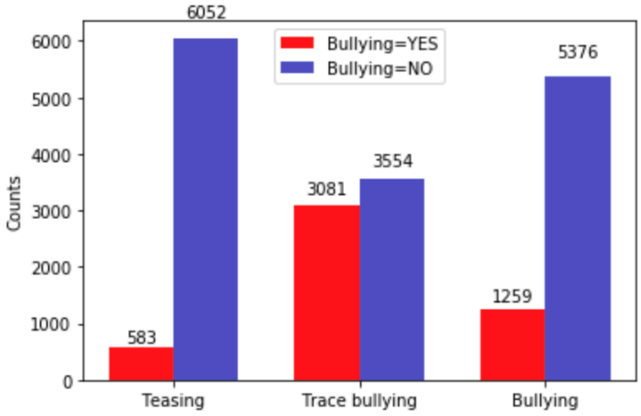


Figure 4 – Results of classifier. Approximately 9% of messages have some form of teasing. In 46%, there is some hint or trace of bullying. About 19% of messages are classified as containing some form of bullying.

[3] https://radimrehurek.com/gensim/models/ldamodel.html

[4] https://github.com/bmabey/pyLDAvis

Table 1 – Topics in the week of Feb 8, 2015 through Fab 15, 2015 using TD-IDF weighted corpus.

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	fuck	geneseo	think	much	apo	rush	girl	anyone	tight	okay
Topic 2	girls	need	life	one	love	never	want	help	shit	fucking
Topic 3	please	everyone	fix	hot	feel	daga	back	coming	sit	face
Topic 4	know	still	night	bids	anyone	waiting	week	day	period	think
Topic 5	worst	times	first	today	want	drunk	sex	wear	feelings	chipotle
Topic 6	night	balls	go	last	new	fuck	freedom	look	girl	going
Topic 7	party	mom	even	things	everyone	found	college	guy	alone	sex
Topic 8	taste	friends	go	want	guy	mom	love	supposed	statesmen	Make
Topic 9	day	sunday	hope	freedom	love	go	watching	want	wish	Shit
Topic 10	food	fucked	long	bed	hate	much	already	everyone	actually	dick

A model with too many topics, will typically have many overlaps, small sized bubbles clustered in one region of the chart [7].

Lastly, the cyberbullying classifier[5] was run on YikYak messages. Results are shown in Figure 4.

4. RESULTS

4.1 Cyberbullying classifier

The YikYak dataset we used was collected in 2015 while YikYak was quite popular. From our analysis, we showed that 1 in every 5 messages contained some form of cyberbullying. This can include physical threats, verbal abuse, teasing, and any other cyberbullying activity.

4.2 Sentiment Analysis

We also observed more than half of the yaks (52%) having a negative sentiment. Our classifier achieved 89% accuracy so there is definitely some error associated with classification. However, this metrics shows us that half of our posts are negative.

4.3 Topic Modeling

Term frequency is not always the most accurate way of determining the importance of a particular word or phrase in social media datasets. This is due to the small nature of each document. However, when doing a direct comparison of the words returned by both the bag-of-words (BOW) approach and the TF-IDF approach, BOW approach outlines more common words, but TF-IDF picks the important words used in each message. Empirically speaking, the TF-IDF provided more nouns and adverbs compared to BOW which provided more parts of language.

Correlating some of the trends, we see that topics of love or companionship emerged during the week of Valentine's Day. Generally though in our dataset, topics of sex, drugs, partying, girls, and food are quite common.

5. CONCLUSION

Our research confirms that if left unmoderated, anonymous social media platforms and micro-communities can quickly degrade and be overrun by cyberbullying and obscene content. Geneseo, which is a quiet little college town, is overrun with themes of drugs, sex, and partying.

Using machine learning, we are able to monitor cyberbullying and hate speech.

6. REFERENCES

[1] Van Hee C, Jacobs G, Emmery C, et al. Automatic detection of cyberbullying in social media text. *PLoS One*. 2018;13(10):e0203794. Published 2018 Oct 8. doi:10.1371/journal.pone.0203794

[2] Constine, Josh. "#1 App YOLO Q&A Is the Snapchat Platform's 1st Hit – TechCrunch." *TechCrunch*, TechCrunch, 8 May 2019, techcrunch.com/2019/05/08/download-yolo-app/.

[3] Caufield, Jane. "Yik Yak Successor Aims to Bring Mobile Chat Back to Campus." *Top Hat*, 10 Oct. 2017, tophat.com/blog/yik-yak-campus-chat/.

[4] Fye, Shaan. "Yik Yak: Why It Exists." *Atlas Business Journal*, 16 Oct. 2016, atlasbusinessjournal.org/yik-yak-greater-implications-upon-society/.

[5] http://research.cs.wisc.edu/bullying/data.html

[5] Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). ACM, New York, NY, USA, 80-88. DOI=http://dx.doi.org/10.1145/1964858.1964870

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022.

[7] "Topic Modeling in Python with Gensim." *Machine Learning Plus*, 4 Dec. 2018, www.machinelearningplus.com/nlp/topic-modeling-gensim-python

[8] Chen, Edwin. "Introduction to Latent Dirichlet Allocation." *Edwin Chen*, blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/.