2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

# Characterizing (Un)moderated Textual Data in Social Systems

Lucas Lima*, Julio C. S. Reis*†, Philipe Melo*, Fabrício Murai*, Fabrício Benevenuto*
*Universidade Federal de Minas Gerais (UFMG), Brazil, †Universidade FUMEC, Brazil
{lucaslima, julio.reis, philipe, murai, fabricio}@dcc.ufmg.br

***Abstract*—Despite the valuable social interactions that online media promote, these systems provide space for speech that would be potentially detrimental to different groups of people. The moderation of content imposed by many social media has motivated the emergence of a new social system for free speech named Gab, which lacks moderation of content. This article characterizes and compares moderated textual data from Twitter with a set of unmoderated data from Gab. In particular, we analyze distinguishing characteristics of moderated and unmoderated content in terms of linguistic features, evaluate hate speech and its different forms in both environments. Our work shows that unmoderated content presents different psycholinguistic features, more negative sentiment and higher toxicity. Our findings support that unmoderated environments may have proportionally more online hate speech. We hope our analysis and findings contribute to the debate about hate speech and benefit systems aiming at deploying hate speech detection approaches.**

***Index Terms*—Social Network, Moderated Content, Unmoderated Content, Hate Speech, Gab, Twitter.**

## I. Introduction

The Web has changed the way our society communicates, giving rise to social platforms where users can share different types of content and freely express themselves through posts containing personal opinions. Unfortunately, with the popularization of this new flavor of communication, toxic behaviors enacted by some users have been gaining prominence through online harassment and hate speech. These platforms have become the stage for numerous cases of online hate speech, a type of discourse that aims at attacking a person or a group on the basis of race, religion, ethnic origin, sexual orientation, disability, or gender [1].

Recently, to prevent the proliferation of toxic content, most online social networks prohibited hate speech in their user policies and enforced this rule by deleting posts and banning users who violate it. Particularly, Twitter, Facebook, and Google (YouTube) have largely increased removals of hate speech content [2] by making available their hate policies so users can actively report content that might violate their policies. Reddit also deleted some communities related to fat-shaming and hate against immigrants [3]. This scenario has motivated the emergence of a new social network system, called Gab. In essence, Gab is very similar to Twitter, but barely moderates any of the content shared by its users. According to Gab guidelines, the website promotes freedom of expression and states that "*the only valid form of censorship is an individual's own choice to opt-out*". They, however, do not allow illegal activity, spam, or form of illegal pornography, promotion of violence and terrorism.

Despite existing recent efforts that attempt to understand the content shared in Gab [4, 5], there is still a need for an understanding regarding the amount and forms of hate speech in an unmoderated system such as Gab. In this article, we provide a diagnostic of hate in the unmoderated content from Gab, by categorizing the different forms of hate speech in that system and comparing it with Twitter, a proxy for a moderated system. Specifically, we identify textual characteristics of a set of unmoderated (or barely moderated) data, in this work represented by Gab, and compare them with characteristics of moderated data, here represented by Twitter. Our study is based on the analysis of $7,794,990$ Gab posts and $9,118,006$ tweets from August 2016 to August 2017. At a high level, our analysis is centered around the following research questions: ***Research Question 1:*** *What are the distinguishing characteristics of moderated content in Twitter and unmoderated content in Gab in terms of linguistic features, sentiment, and toxicity?* ***Research Question 2:*** *What are the most common types of hate in an unmoderated and moderated environment?*

To answer our first research question we quantify the linguistic differences across moderated and unmoderated social systems by applying established language processing techniques. Our analysis shows that content in Gab and Twitter have different linguistics patterns, with higher toxicity and a more negative overall sentiment score in the unmoderated Gab content. Additionally, we find that, in general, Gab has more hate posts than Twitter. We show that Gender and Class types of hate are more frequent on Twitter, whereas Disability, Ethnicity, Sexual Orientation, Religion, and Nationality types tend to appear proportionality more in Gab.

These findings highlight the importance of creating moderation policies as an effort to fight online hate speech in social systems, and also point out possible points for improvement in the design of content policies for social media systems. Additionally, our findings suggest that the unmoderated content found in Gab might be an appropriate data source for the development of learning approaches to detect hate speech. Thus, as a final contribution, we make our hate-labeled Gab posts available for the research community[1] can foster the development of future hate speech detection systems.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

[1] https://github.com/lhenriquecl/unmoderated-hate-dataset

## II. Related Work

### A. Social Media Content Moderation

The ongoing discussion on content moderation in an online environment led Internet companies to reflect about the undesirable attention their sites can attract and the consequences of it[2], thus the first steps towards regulating and moderating content have already been taken. There is a lot of debate on how social media platforms moderate their own content, and how their moderation policies are shaped. Twitter and YouTube, for instance, make available their hate policies[3,4] so users can actively report content that might violate their policies. Our work contributes to this discussion since we quantify the differences between moderated and unmoderated text data as an effort to shed light on the importance of creating different methodologies and policies to outline the boundaries of hate in social media.

Besides, some studies have focused on the understanding of systems that lack moderation of content. Finkelstein et al. [6] focus on making an extensive analysis of antisemitism in 4chan's and Gab. Their results provide a quantitative data-driven framework for understanding this form of offensive content. Zannettou et. al [5] and Lima et. al [4] present the first characterization studies on Gab, analyzing network structure, users, and posts. Our effort is complementary to these works as we compare the textual data shared from an unmoderated system like Gab with a moderated one as Twitter and deeply investigate the types of hate in both social systems.

### B. Online Manifestations of Hate Speech

A vast number of studies were conducted to provide a better understanding of hate speech on the Internet. Silva et al. and Mondal et al. [7, 8] provide a deeper understanding of the hateful messages exchanged in social networks, studying who are the most common targets of hate speech in these systems. Salminen [9] create a taxonomy and use it to investigate how different features and algorithms can influence the results of the hatefulness classification of text using several machine learning methods. Chandrasekharan et al. [10] characterize two banned communities from *Reddit*, one about fat shaming and the other related to hate against immigrants in the US, and proposed a lexicon for hate speech detection. Our effort is complementary to these studies, as we quantify the amount of hate shared in a moderated and in an unmoderated environment and highlight the different types of hate exchanged in these social systems, elucidating the importance of content moderation to fight hate on the Internet.

Several other efforts have attempted to provide detection approaches for hate speech [11, 12, 13, 14]. As a final contribution, we make the hate-labeled Gab dataset available to the research community. We hope our efforts help other hate speech studies on the creation of better methods for identifying hate on social media.

[2]https://www.nytimes.com/2010/07/19/technology/19screen.html
[3]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[4]https://support.google.com/youtube/answer/2801939?hl=en

## III. Datasets and Methods

### A. Datasets

Our Gab dataset comprises posts from users crawled following a Breadth-First Search (BFS) scheme on the graph of followers and friends. We used as seeds users who authored posts listed by categories on the Gab main page. We implemented a distributed and parallel crawler that ran in August 2017. In total, our Gab dataset comprises $12,829,976$ posts, obtained from $171,920$ users (the estimated number of users in August 2017 was 225 thousand [15]).

The Twitter dataset contains English posts randomly selected from the Twitter $1\%$ Streaming API. For consolidating our dataset and keep data consistency, we consider only random tweets published in the same period as Gab posts, which gives us also $12,829,976$ tweets. After preprocessing and removing duplicated posts in both datasets, we have a total of **7,794,990 Gab posts** and **9,118,006 tweets**. These are the final sets of posts for each media that are going to be further analyzed in this work.

### B. Language Processing Methods

Next, we describe the methods used in this work to perform language characterization.

*1) Linguistic Analysis:* One of our goals is to understand the distinguishing linguistic characteristics of posts on Gab and Twitter and contrast them. Thus, we use the 2015 version of the Linguistic Inquiry and Word Count (LIWC) [16] to extract and analyze the distribution of psycholinguistic elements posts of both media. LIWC is a psycholinguistic lexicon system that categorizes words into psychologically meaningful groups all of which form the set of LIWC attributes, and has been widely used for several different tasks [17, 18, 19, 20].

*2) Sentiment Analysis:* We perform sentiment analysis on Gab and Twitter posts as a complementary effort to characterize the differences between moderated and unmoderated accounts. We use an established opinion mining method to measure sentiment score on our messages: the SentiStrength [21], which has shown to be an effective tool for sentiment analysis in social media posts [22]. We apply the standard English version of it to quantify positive $P \in \{+1, \ldots, +5\}$ and negative $N \in \{-5, \ldots, -1\}$ sentiments in each post, as well as their overall sentiment score, which is given by the difference between $P$ and absolute $N$ values for a post.

*3) Toxicity Analysis:* We measure the toxicity of posts with the Perspective API, created by Jigsaw and Google's Counter Abuse Technology team, also as a complementary analysis to elucidate the difference between moderated and unmoderated content. This score measures how "*toxic*" a message can be perceived by a user. Toxic messages are defined as *a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion*. The value does not represent a degree of "toxic severity" of a particular message, but instead the probability that someone perceives that message as toxic. Scores range from 0 to 1, where scores closer to 1 indicate that posts are likely to be perceived as toxic.

Fig. 1: Flowchart representing the method for identifying hate posts from Gab and Twitter posts.

TABLE I: Lexicon of categorized hate terms.

| | |
|---|---|
| **Ethnicity** | bamboo coon, boojie, camel fu**er, house ni**er, moon cricket, ni**er, plastic paddy, raghead, sideways pu**y, spic, trailer park trash, trailer trash, wetback, whi**er, white ni**er, white trash, wi**er, zionazi |
| **Class** | bitter clinger, boojie, redneck, rube, trailer park trash, trailer trash, white trash, yo**o |
| **Disability** | retard, retarded |
| **Nationality** | bamboo coon, camel fu**er, chinaman, limey, plastic paddy, sideways pu**y, soup taker, surrender monkey, whi**er, white ni**er, wi**er, zionazi |
| **Religion** | camel fu**er, muzzie, soup taker, zionazi |
| **Gender** | bint, cu*t, d*ke, t*at |
| **Sexual Orientation** | d*ke, fa**ot |

### C. Assessing Hate Speech

ElSherief et al. [23] present a semi-automated classification approach for the analysis of directed explicit hate speech which relies on keyword-based methods and on the Perspective API. The authors validate their methodology by incorporating human judgment using Crowdflower, concluding that their final hate speech dataset is reliable and has minimal noise. The method to detect hate implemented in this work is inspired by the referred work and it is similar to it with minor changes.

Figure 1 illustrates the method for identifying hate posts implemented in this work. First, both datasets go through a pipeline where the initial step is to query the Perspective API using each post as input. Besides the toxicity score, we gather the *attack on commenter* score of posts, which measures direct and personal offense or injury to another user participating in the discussion. Next, we filter posts which have toxicity score higher than 0.8 and attack on commenter higher than 0.5 (these thresholds were defined by ElSherief et al. [23] so as to yield a high quality dataset). Finally, we check whether these filtered posts contain at least one hate word, and, if so, we assume these are hate posts.

The list of hate words is also obtained from the study of ElSherief et al. [23]. Table I shows the terms[5] from the categorized lexicon which are currently in Hatebase and are associated with at least one type of hate. Following the aforementioned methodology, **9,554** (0.12%) Gab posts and **2,392** (0.03%) tweets are labeled as **hate** and are going to be further explored on our hate speech analysis.

## IV. RQ1: Distinguishing Characteristics of Moderated and Unmoderated Content

### A. Linguistic Features

We analyze linguistic differences of moderated and unmoderated content by computing the distributions values for each LIWC attribute in both sets of posts. We aggregate these attributes into four distinct dimensions (*Standard Linguistic Dimensions*, *Personal Concerns*, *Spoken Categories*, and *Psychological Processes*) following Pappas et al. [24] who made this arrangement available[6]. We start by investigating the volume of posts from both social media containing words in each LIWC dimension, as shown in Figure 2.

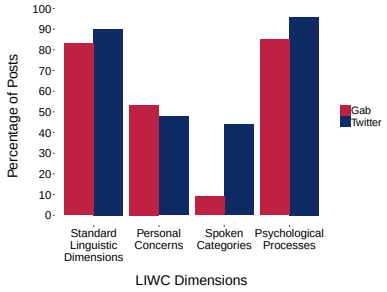



Fig. 2: Percentage of Gab and Twitter posts which contain at least one word or token per LIWC dimension.

More than 80% of Gab and Twitter posts contain at least one term of either the Standard Linguistic Dimensions or the Psychological Processes dimensions. Nearly 50% of posts from both social media contain words of Personal Concerns. Interestingly, for the Spoken Categories, 43.9% of Twitter posts contain at least one word of this dimension, whereas only 9.1% of Gab posts contain at least one of the referred words. This difference might be due to the characteristics of the audience and posts of the Gab network, which has a strong political bias where users tend to share a larger number of news and politics related posts [4], whereas Twitter has traits of behavior that may encourage informal communication [25]. We compare the distributions of each LIWC attribute for both Gab and Twitter by running the Kolmogorov-Smirnov (KS) test [26]. We find significant statistical difference (p-value $< 0.05$) for all the distributions, indicating that moderated and unmoderated posts have different psycholinguistic features.

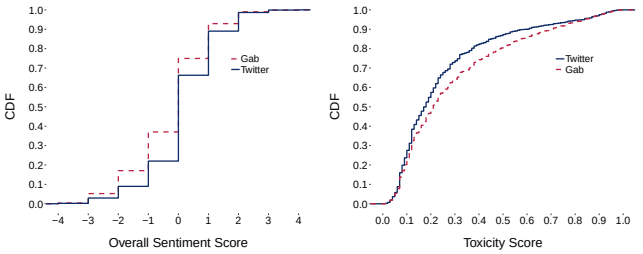
### B. Sentiment Analysis and Toxicity

Next, we analyze the differences of sentiment and toxicity for moderated and unmoderated content. Figure 3 shows the Cumulative Distribution Function (CDF) for the (a) overall sentiment score, calculated as the difference between the positive and the absolute value of negative scores given by SentiStrength ($P - |N|$), and (b) toxicity score for both Gab and Twitter. First, we compare these distributions using the KS test. For each metric, we find a significant statistical difference between the distributions, i.e., posts from moderated and unmoderated social media are statistically different in terms of sentiment and toxicity scores.

We notice that unmoderated publications tend to be more negative and to have higher toxicity than moderated ones. For the overall sentiment scores, Figure 3a shows that nearly 37% of unmoderated posts have negative overall sentiment, i.e., have an absolute negative score greater than a positive score,

[5]Wherever present, the '*' has been inserted by us, in order to lessen the impact that the offensive terms may inflict on some people, and was not part of the original word or text

[6]https://lit.eecs.umich.edu/geoliwc/liwc_dictionary.html




(a) (b)

Fig. 3: CDF for (a) overall sentiment and (b) toxicity scores.

TABLE II: Manual evaluation. Each triple shows the number of posts agreed as hate, without agreement and agreed as non-hate, respectively.

| | Gab | Twitter |
|---|---|---|
| **Labeled by framework as Hate** | (93, 4, 3) | (90, 6, 4) |
| **Labeled by framework as Non-Hate** | (5, 4, 91) | (0, 3, 97) |

whereas only 22% of moderated posts have negative overall sentiment. Furthermore, Figure 3b shows that 19.4% of Gab posts have toxicity scores higher than 0.5, whereas this percentage is 13% for Twitter posts, indicating that unmoderated posts tend to be perceived as toxic more often than moderated posts. The fraction of posts that have toxicity above 0.8 is higher in Gab than on Twitter, as there are 6.5% of such posts on Gab and 5.5% of tweets in all posts. These results suggest that Gab is a more toxic social network than Twitter. One probable explanation for this is the lack of moderation on Gab, allowing harmful speech in this network to fester unchecked.

## V. RQ2: Different Types of Hate Across Moderated and Unmoderated Environments

### A. Manual Validation

We first evaluate the quality of our hate set by manually annotating a random sample of Gab and Twitter posts. We take 100 random posts that were marked as hate/non-hate by the framework described in Section III-C. Two authors of this work independently annotated these posts by hand as hate or non-hate according to their perspectives. We hide the previously assigned labels from the annotators. Table II shows the results of our manual evaluation for the given samples. We notice the occurrence of few false positive ($\leq 4\%$) and false negative ($\leq 5\%$) considering the agreement of the annotators.

The annotators have agreed on the label of 383 posts, resulting in a Cohen's kappa coefficient $\kappa = 0.92$. Assuming the labels created by the annotators of these 383 posts as the correct ones, the method to identify hate presented in this work was able to correctly classify 371 labels: accuracy 96.87%, precision 96.32%, and recall 97.34%. These results reinforce the quality of the method to identify hate posts with minimal noise.

### B. Types of Hate

We associate one hate post with different types of hate, according to the types shown in Table I: *ethnicity*, *class*, *disability*, *nationality*, *religion*, *gender*, and *sexual orientation*.

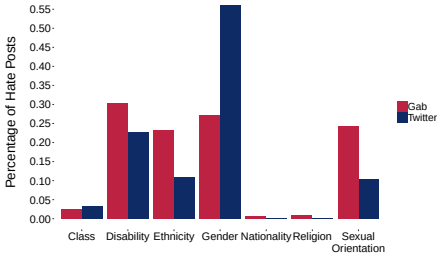



Fig. 4: Percentage of posts associated with each type of hate.

TABLE III: Top 5 most frequent hate terms in Gab and Twitter hate posts.

| Gab | | Twitter | |
|---|---|---|---|
| **term (%)** | **category** | **term (%)** | **category** |
| retarded (30.25) | disability | c*nt (45.31) | gender |
| fa**ot (23.96) | sexual orientation | retarded (22.74) | disability |
| c*nt (22.10) | gender | t*at (9.65) | gender |
| ni**er (21.3) | ethnicity | fa**ot (9.53) | sexual orientation |
| t*at (4.68) | gender | ni**er (8.52) | ethnicity |

One hate post can be associated with up to seven types of hate (or none). Being associated with one type of hate does not necessarily imply that one post is being hateful towards that particular group, but rather it makes use of terms associated with that type of hate to write a comment that is perceived as rude, disrespectful, or unreasonable by different groups.

*1) Frequent types of hate:* Considering our hate set, Figure 4 shows the percentage of Gab and Twitter posts associated with each type of hate. Our findings show that Gab and Twitter posts are predominantly associated with disability and gender types of hate. It is interesting to notice that 56.06% of hate tweets are associated with gender, followed by 22.74% of hate tweets being related to disability, a difference of over 30%.

For Gab, this difference is smaller, as disability is associated with 30.25% and is followed by gender, with 27.04%. Gab has also a large number of hate posts associated with sexual orientation and ethnicity (over 20% each). These results suggest that an environment which lacks moderation of content like Gab is more prone to the dissemination of hate speech of many types than a moderated one, which still needs to improve their hate speech policies and methods in order to avoid hate speech towards specific groups of people.

*2) Frequent hate terms:* Table III shows the 5 most frequent hate terms in Gab and Twitter hate sets. The complete rankings have significant Kendall rank correlation coefficient (0.76). We observe that all gender related hate terms appear proportionally more on Twitter than Gab, which helps explaining the larger number of gender related hate posts observed in Twitter in comparison with Gab on Figure 4. The term *c*nt* appears in more than 45% of Twitter hate posts whereas this number for Gab drops to near 22%. *T*at* and *d*ke* are the others gender related terms on the top 10 which appear proportionality more on Twitter hate posts than Gab's, corroborating our findings on the analysis of linguistic differences between these networks.

### C. Potential Limitations

Our Twitter dataset is shaped by the limitations of getting a sample from all Twitter with the Streaming API [27]. Moreover, hate speech classification is inherently difficult, as there is no universal definition for it and many important variables, such as context, are not easily measured. Our method to detect hate relies, as a first step, on an external API which might also lead to inaccurate toxicity scores for social media posts. Hosseini et al. [28] have shown that subtle changes on highly toxic sentences may assign significantly lower scores to them, which may indicate that many posts could not be classified as hateful in our work. Nevertheless, we have showed that by building on previous our approach can accurately identify many forms of hate speech.

## VI. Conclusion

In this work, we provide an in-depth quantitative analysis of moderated data from Twitter and unmoderated data from Gab, a social network that received several criticisms regarding the content shared on it. We perform linguistic analysis and conduct an investigation of hateful posts and the various shades of hate that are displayed in this right-leaning echo chamber. Our analysis on Gab, put into perspective with Twitter, showed that the unmoderated posts on Gab present more negative sentiment, higher toxicity, and different psycholinguistic features. Our findings support that unmoderated environments may have proportionally more hate speech. Furthermore, we categorize hate speech and its different forms in both environments, unveiling a highly toxic discourse in Gab in the many forms that hate speech can manifest itself.

Our work makes an important step towards the development of automated hate speech detectors. We believe that an unmoderated hate dataset[7], as the one analyzed, can help the development of hate speech detection approaches in future works. For this reason, our final contribution consists of making our hate-labeled Gab posts available to the community.

## Acknowledgments

This work was partially supported by the MPMG, project Analytical Capabilities, CNPq, CAPES, and Fapemig.

## References

[1] N. Johnson, R. Leahy, N. J. Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty, "Hidden resilience and adaptive dynamics of the global online hate ecology," *Nature*, 2019.

[2] A. Macdonald and J. Fioretti, "Social media firms have increased removals of online hate speech: EU," shorturl.at/rAJKO, 2017.

[3] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *PACM HCI*, 2017.

[4] L. Lima, J. C. S. Reis, P. Melo, F. Murai, L. A. Araujo, P. Vikatos, and F. Benevenuto, "Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system," in *ASONAM*, 2018.

[5] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn, "What is Gab: A Bastion of Free Speech or an Alt-right Echo Chamber," in *WWW*, 2018.

[6] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn, "A quantitative approach to understanding online antisemitism," *arXiv preprint arXiv:1809.01644*, 2018.

[7] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in *ICWSM*, 2016.

[8] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *HT*, 2017.

[9] J. Salminen, H. Almerekhi, M. Milenkovic, S.-g. Jung, J. An, H. Kwak, and B. J. Jansen, "Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media." in *ICWSM*, 2018.

[10] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *PACM HCI*, November 2017.

[11] J. Bartlett, J. Reffin, N. Rumball, and S. Williamson, "Anti-social media," *Demos*, pp. 1–51, 2014.

[12] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *IJMUE*, vol. 10, no. 4, pp. 215–230, 2015.

[13] S. Agarwal and A. Sureka, "Using knn and svm based one-class classifier for detecting online radicalization on twitter," in *ICDCIT*. Springer, 2015, pp. 431–442.

[14] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *LSM*. ACL, 2012.

[15] K. Flynn, "Tech is cracking down on hate speech—but it's still thriving on Gab and 4chan," shorturl.at/grsBW, 2017.

[16] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *JLS*, 2010.

[17] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, "Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ*, 2016.

[18] J. C. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised learning for fake news detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.

[19] D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi, "The many shades of anonymity: Characterizing anonymous social media content," in *ICWSM*, 2015.

[20] G. Resende, P. Melo, J. C. S. Reis, M. Vasconcelos, J. Almeida, and F. Benevenuto, "Analyzing textual (mis)information shared in whatsapp groups," in *WebSci*, 2019.

[21] M. Thelwall, "Heart and soul: Sentiment strength detection in the social web with sentistrength," *CyberEmotions*, 2013.

[22] A. Abbasi, A. Hassan, and M. Dhar, "Benchmarking twitter sentiment analysis tools," in *LREC*, 2014.

[23] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in *ICWSM*, 2018.

[24] K. Pappas, S. Wilson, and R. Mihalcea, "Stateology: State-level interactive charting of language, feelings, and values," *arXiv preprint arXiv:1612.06685*, 2016.

[25] D. Zhao and M. B. Rosson, "How and why people twitter: the role that micro-blogging plays in informal communication at work," in *ACM GROUP*, 2009.

[26] F. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *JASA*, 1951.

[27] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *ICWSM*, 2013.

[28] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv preprint arXiv:1702.08138*, 2017.

[7] https://github.com/lhenriquecl/unmoderated-hate-dataset