Detection of Abusive Language from Tweets in Social Networks

Ms. Mohini S. Dadhe
Dept.of Computer Engineering
BapuraoDeshmukh College of Engineering, Sevagram,
Wardha

Ms. Pranali S. Masidkar
Dept.of Computer Engineering
BapuraoDeshmukh College of Engineering, Sevagram,
Wardha

Ms. Vishanka Vaidya
Dept.of Computer Engineering
BapuraoDeshmukh College of Engineering, Sevagram,
Wardha

Prof. Priyanka A. Jalan
Dept of Computer Engineering
BapuraoDeshmukh College of Engineering

ABSTRACT: Detection of abusive language in user generated online con-tent has become an issue of increasing importance in recent years. Most current commercial methods make use of black-lists and regular expressions, however these measures fall short when contending with more subtle, less ham-fisted ex-samples of hate speech. In this work, we develop a machine learning based method to detect hate speech on online user comments from two domains which outperforms a state-of-the-art deep learning approach. We also develop a corpus of user comments annotated for abusive language, the first of its kind. Finally, we use our detection tool to analyze abusive language over time and in different settings to further enhance our knowledge of this behavior.

Keywords*: Density Based Clustering Algorithm*

*****

I. Introduction:

Anytime one engages online, whether on message board forums, comments, or social media, there is always a serious risk that he or she may be the target of ridicule and even harassment. Words and sentences such as kill yourself asshole or they should all burn in hell for what they've done are unfortunately not uncommon online and can have a profound impact on the civility of a community or a user's experience. To combat abusive language, many internet companies have standards and guidelines that users must adhere to and employ human editors, in conjunction with systems which use regular

expressions and blacklist, to catch bad language and thus remove a post. As people increasingly communicate online, the need for high quality automated abusive language classifiers becomes much more profound. Recent cases highlight the impact of hurtful language in online communities, as well as on major corporations.

For example, in 2013, Facebook came under fire for hosting pages which were hateful against women such as Violently raping your friend just for laughs and Kicking your girlfriend in the fanny because she won't make you a sandwich. 1 Within days, a petition was started which amassed over 200,000 supporters, and several major companies either pulled or threatened to pull their ads from Facebook since they were inadvertently places on these pages. Facebook is not the only company that contends with these issues; any company which hosts user generated content will have a moderation issue. This shows the large impact hateful language can have on a community as well as a major company. At the more individual level, when actor Robin Williams passed away, his daughter Zelda posted a memoriam to her late father and was immediately bullied on Twitter and Instagram and eventually deleted all of her online accounts. This harrassment prompted Twitter to review and revise its hate speech guidelines.2 While automatically detecting abusive language online is an important topic and task, the prior art has not been very unified, thus slowing progress. Past research has spanned different fields ranging from Natural Language Processing (DBCA) to Web Sciences to Artificial Intelligence, meaning that several similar methods were published in the last three years. Additionally, abusive language can be a bit of a catchall term. There are some studies, [14] which focus on detecting profanity, and others, such as [18] which focus on hate speech directed to a particular ethnic group. To further complicate matters, to date there has been no de facto testing set with which to compare methods.

In this paper we aim to develop a state-of-the-art method for detecting abusive language in user comments, while also addressing the above deficiencies in the field. Specifically, this paper has the following contributions:

- We develop a supervised classification methodology with DBCA features to outperform a deep learning approach. We use and adapt several of the features used in prior art in an effort to see how they perform on the same data set. We also extend this feature set with features derived from distributional semantics techniques.

- We make public a new data set of several thousand user comments collected from different domains. This set includes three judgments per comment and for comments which are labeled as abusive, a more fine-grained classification on how each is abusive.

• Prior work has evaluated on a fixed, static data set. However, given the issues with language changing over time and also with users trying to cleverly evade keyword based approaches, we perform several analyses of how models trained on different types and sizes of data perform over the span of one year, across two different domains. To our knowledge, this is the first longitudinal study of a computational approach to abusive language detection.

II. Related Work

Most prior work in the area of abusive language detection has actually been spread across several overlapping fields. This can cause some confusion as different works may tackle specific aspects of abusive language, define the term differently, or apply it to specific online domains only (Twitter, online forums, etc.). To further complicate comparison between approaches, nearly all previous work uses different evaluation sets. One of the contributions of this paper is to provide a public dataset in order to better move the field forward. One of the first works to address abusive language was [21] which used a supervised classification technique in conjunction with n-gram, manually developed regular expression patterns, contextual features which take into account the abusiveness of previous sentences. As most basic approaches make use of predefined blacklists, [15] noted that some blacklist words might not be abusive in the proper context. In their work they showed an improvement in profanity detection by making use of lists as well as an edit distance metric. The latter allowed them to catch such un-normalized terms as @ss or sh1t. Another contribution of the work was that they were the first to use crowdsourcing to annotate abusive language. In their task, they used Amazon Mechanical Turk workers to label 6,500 internet comments as abusive or not abusive. They only used comments in which a majority of the turkers agreed on the label. 9% of the comments were deemed as carrying profane words. In our work, we also make use of crowdsourcing to curate a corpus of several thousand internet comments. The main differences are that we do not limit the task to just profanity and also have the workers annotate for other types of hate speech and abusive language. In addition, we are making this dataset public. [3] was one of the first to use a combination of lexical and parser features to detect offensive language in youtube comments to shield adolescents. While they do note that they do not have a strict definition of offensive language in mind, their tool can be tuned by the use of a threshold which can be set by parents or teachers so online material can be filtered out before it appears on a web browser. The work takes a supervised classification approach using Support Vector Machines (SVMs) with features including n-grams, automatically derived blacklists, manually developed regular expressions and dependency parse features. They achieve a performance on the task of inflammatory sentence detection of precision of 98.24% and recall of 94.34%.

One difference between our work and this one is that they attempt to spell correct and normalize noisy text before feature extraction. We believe that this noise is a potentially good signal for abuse detection and thus have features to capture different types of noise. Our work also makes use of dependency features, though with a much broader set of tuples than [3]. [18] provide the most comprehensive investigation of hate speech (hateful language directed towards a minority or disadvantaged group) to date, with working definitions and an annotation task. Here their focus was less on abusive language and more specifically on anti-semitic hate. First, they manually annotated a corpus of websites and user comments, with Fleiss kappa interlabeler agreement at 0.63. Next, they adopted a related approach to the aforementioned supervised classification methods by first targeting certain words that could either be hateful or not, and then using Word Sense Disambiguation techniques [20] to determine the polarity of the word. Their method performs at 0.63 F-score. To our knowledge, this is the only work to target hate speech and the only one to have done a rigorous annotation of data, though the set could not be made public. We build on their work by crowdsourcing the annotation of a data set of user comments, categorizing each comment as abuse, profanity, and/or hate speech. This set will be made public.

Finally, [5] use a paragraph2vec approach adopted from [8] to classify language on user comments as abusive or clean. Their approach outperformed a bag-of-words (BOW) implementation (0.8007 to 0.7889 AUC). In our work, we use a more sophisticated algorithm to learn the representation of comments as low-dimensional dense vectors. Moreover, our representation is learned using only unigrams in order to compliment other relevant features. In our work, we aim for a method that is efficient and flexible but also operates at a high accuracy by combining different light-weight features. We include an evaluation using their data to directly compare our system but also experiment with their approach as additional features in our methodology.

III. METHODOLOGY

For our work we employ a supervised classification method which uses DBCA features which measure different aspects of the user comment. Specifically, we use the VowpalWabbit's regression model5 in its standard setting with a bit rate of 28. We base our DBCA features on prior work in sentiment [9], text normalization [1] among others.

Our features can be divided into four classes: N-grams, Linguistic, Syntactic and Distributional Semantics. For the first three features, we do some mild pre-processing to transform some of the noise found in the data which could impact the number of sparse features in the model. Example transformations include normalizing numbers, replacing very long unknown words with the same token, replacing repeated punctuation with the same token, etc. For the fourth feature class, we did none of the above normalization.

Density Based Algorithm

- Density Based Algorithm is also known as Denstream Algorithm that work on density of the element.
- Two important parameters are required ε (epsilon) and Minimum points.
- ε defines the radius of neighborhood around a point x and Minimum points defines the minimum number of neighborhood within x.
- Includes Micro clusters such as Corepoint,border point,Noise point.

- For every point p in a cluster C there is a point q ∈ C, so that ..,p is inside of the Eps-neighborhood of q And $N_{Eps}(q)$ contains at least MinPts points.

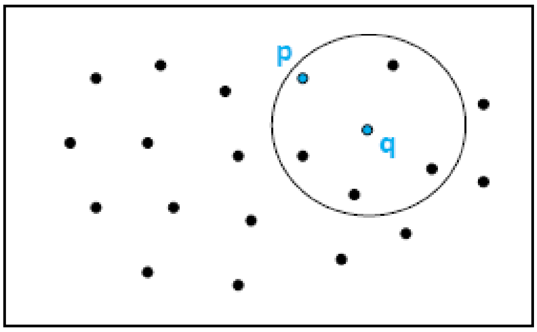


Figure 1. Neighbourhood Points

- A point p is density-connected to a point q with regard to the parameters Eps and MinPts if there is a point v such that both p and q are density-reachable from v.

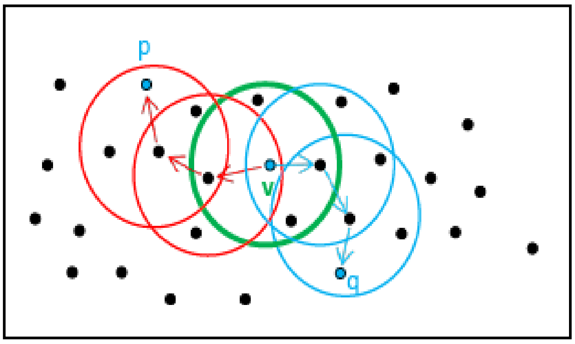
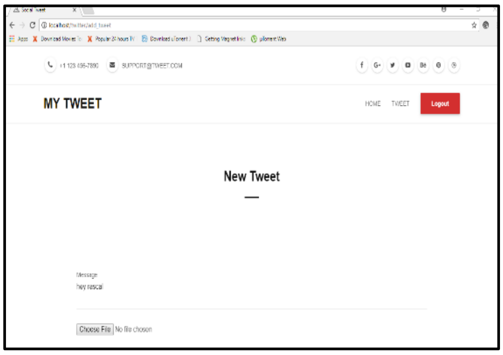


Figure 2. Reachability from p to q

- For all p, q ∈ D: If p ∈ C and q is density-reachable from p with regard to the parameters Eps and MinPts, then q ∈ C.

- For all p, q ∈ C: The point p is density-connected to q with regard to the parameters Eps and MinPts.

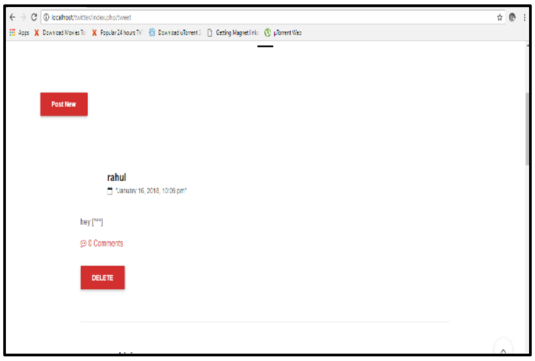
IV. Experimental Results



New Tweet

Message
hey rascal

Snapshot 1: Page to Tweet



rahul
"January 16, 2018, 10:09 pm"

hey ["*"]

Snapshot 2: Bad words in tweet replaced by *

V. Conclusion

The copy-move forgery detection is one of the emerging problems in the field of Text and multimedia systems. In the last decade many forgery detection techniques have been proposed. An attempt is made to bring in various potential algorithms that signify improvement in bad and malicious text detection techniques. The techniques which have been developed till now are mostly cable of detecting the forgery and only a few can localize the tampered area. There are many drawbacks with the presently available technologies. Firstly all systems require human interpretation and thus cannot be automated.

References

[1]. S. Brody and N. Diakopoulos“ using word lengthening to detect sentiment in microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 562–570, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[2]. M. D. Buhrmester, T. Kwang, and S. D. Gosling. Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6(1):3–5, Jan 2011.

[3]. Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pages 71–80. IEEE, 2012.

[4]. N. Djuric, H. Wu, V. Radosavljevic, M. Grbovic, and N. Bhamidipati. Hierarchical neural language models for joint representation of streaming documents and their content. In International World Wide Web Conference (WWW), 2015.

[5]. N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In Proceedings of International World Wide Web Conference (WWW), 2015.

[6]. M. Faruqui and C. Dyer. Non-distributional word vector representations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 464–469, Beijing, China, July 2015. Association for Computational Linguistics.

[7]. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. National Bureau of Economic Research Cambridge, Mass., USA, 2010.

[8]. Q. Le and T. Mikolov. Distributed representations of sentences and documents. In T. Jebara and E. P. Xing, editors, Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1188–1196. JMLR Workshop and Conference Proceedings, 2014.

[9]. B. Liu. Sentiment Analysis and Opinion Mining. Morgan Claypool Publishers, 2012.

[10]. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.

[11]. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013.

[12]. G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. Judgment and Decision Making, 5(5):411–419, 2010.

[13]. E. Pitler and A. Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In Proceedings of the ACL-IJCDBCA 2009 Conference Short Papers, pages 13–16, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[14]. S. Sood, J. Antin, and E. Churchill. Profanity use in online communities. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1481–1490. ACM, 2012.

[15]. S. O. Sood, J. Antin, and E. F. Churchill. Using crowdsourcing to improve profanity detection. In AAAI Spring Symposium: Wisdom of the Crowd, 2012.

[16]. M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers to non-factoid questions from web collections. Computational Linguistics, 37:351–383, 2011.

[17]. S. Suri and D. J. Watts. Cooperation and contagion in web-based, networked public goods experiments. PloS One, 6(3), 2011.

[18]. W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, pages 19–26, Montr´eal, Canada, June 2012. Association for Computational Linguistics.

[19]. B. Yang, W. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. CoRR, abs/1412.6575, 2014.

[20]. D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196. Association for Computational Linguistics, 1995.

[21]. D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2:1–7, 2009.