General Review of The Use of Offensive Language in Social Media

Suad Sahib Alag Al Tamimi
Suez Canal University
Email suaads.t2000@gmail.com

مراجعة عامة للغة المستهجنة في وسائل التواصل الإجتماعي

سعاد صاحب التميمي
جامعة قناة السويس

المستخلص

تهدف هذه الدراسة إلى وصف اللغة المستهجة بشكل عام ومناقشة استخدام تلك اللغة في وسائل التواصل الاجتماعي على وجه الخصوص. قدمت العديد من الدراسات السابقة مناقشات للغة بشكل عام ولكن التركيز على اللغة المستهجنة لم يكن بالقدر الكافي. وقد توصلت الباحثة من مراجعة الدراسات السابقة إلى جمع المعلومات التي تتعلق باستخدام اللغة المستهجنة. ومن ثم، يعد استخدام اللغة المستهجنة مشكلة شائعة للسلوك المسيء على شبكات التواصل الاجتماعي عبر الإنترنت. وقد كشفت الباحثة ايضا ان بعض الدراسات قد استخدمت نماذج التعلم الآلي لكشف اللغة المستهجنة والسلوك المسيء في وسائل التواصل الإجتماعي.

الكلمات المفتاحية: اللغة – اللغة المستهجنة – وسائل التواصل الإجتماعي – الخطاب العدائي – الألفاظ العامية.

Abstract

This study aims to describe the offensive language in general and discuss the use of offensive language in social media in order to give a general review for the use of offensive language. However, many previous studies have discussed the language in general but the focus on offensive language is rare. Based on the previous studies, the researcher found from that the use of offensive language is a common problem of abusive behavior on online social media networks.

Various studies have analyzed this problem by using different machine learning models to detect abusive behavior.

Key Words: Language, offensive language, social media, aggressive speech, slang

1. Introduction

In today's world, everyone relies on social media sites like Facebook, Twitter, and Instagram to keep up with current events, connect with friends and family, and publish their views. However, there are some dangers and difficulties that are associated with the use of social media. Aggressive language in user communications has become a major issue with the rise in popularity of social media and online discussion forums. Very often, people tend to be more cavalier and careless with their words when they are communicating virtually as they feel that they are safe as long as there is no physical communication with others. Offensive language has become one of the major issues that attract interests of researchers which they study human interaction in online social media. Scholars focus on the fact that social media aggression is a serious issue that disproportionately impacts the use of language (Hamm et al., 2015, Kowalski and Limber, 2013).

Social media can be a breeding ground for aggressive, provocative, and hateful speech that targets a wide range of social issues, including immigration, racism, gender, weight, and religion. "Many forms of hate speech involve direct insults, but there are also cases where the intended target of the message is not directly named and the message nonetheless contains a demeaning or humiliating tone or message" (Waseem et al., 2017). Yet, the purpose of this

study is to provide an overview of the prevalence of abusive language on social media.

2. Offensive Language

In online debates and social networks, offensive language is frequently used. Offensive language includes swearing, racial epithets, and hate speech (Sigurbergs-son & Derczynski, 2020). The phrase "hate speech" describes hostile or derogatory comments made about a person or group based on that person's or group's defining characteristic. Inflammatory language has the potential to incite actual acts of hate violence. Due to the massive amount of content published, automatic moderation is necessary to identify inappropriate material in social media.

Offensive language includes slang, slurs, and other words and phrases that are typically viewed as offensive or disrespectful. The term "low register" is sometimes used to describe offensive language. It generally refers to "a particular choice of diction or vocabulary seen as acceptable for a certain topic or social circumstance" (Murray et al., 1884).

2.1. Types of Offensive Language

Within the offensive language category, the following subcategories can be found:

1) Swear words which include both literal and figurative swears that are meant to hurt or belittle another person (Wajnryb, 2005). This larger category can be broken down into subcategories. "Cursing calls forth a superior being; it is more ritualistic and intentionally conveyed [...] and it needs not involve harsh words" (Wajn- ryb, 2005: 20). One example of a curse word with an insulting undertone

is "this is a horrible piece of work" (Wajnryb, 2005: 17). The final classification describes the insulting set. According to Wajnryb (2005: 19), swear words and taunts like "fuck you, maniac" are intertwined in everyday speech. So, we may classify derogatory language directed at another as an insult. By "oath," we imply either a formal vow (Hughes, 2006) or, more to the point, a "loose metaphoric curse," as in "He whispered an oath as the hammer impacted his finger" (Wajnryb, 2005: 20).

2) Expletives which convey strong emotions like anger, frustration, delight, and surprise through the employment of powerful, emotionally laden swear words or phrases (Wajnryb, 2005: 18–19). Expletives that are not aimed at a specific individual, such as the exclamations "shit!", "fuck," and "fucking hell!" are used to express the speaker's displeasure with a given situation.

3) Invectives, a more refined variant of the insult, are sometimes used in formal contexts (Wajnryb, 2005: 20). Because this category bypasses the customary lexicon in favor of sarcasm, humor, and wordplay, it is more of an insult than a swear word. It allows the speaker to be dismissive of the other without really using rude language, much with the phrase "you dazzling wit" (Wajnryb, 2005: 20).

3. The Use of Offensive Language in Social Media

Offensive language is a widespread kind of cyberbullying on social networking websites. Machine learning models have been used to try and identify abusive conduct (Xiang et al., 2012; Warner and Hirschberg, 2012; Wang et al., 2014; Nobata et al., 2016; Burnap

and Williams, 2015; Davidson et al., 2017; Founta et al., 2018). The underlying premise of these works is that it is sufficient to filter out the full objectionable post. A user who is consuming online content, however, may not wish for a completely filtered out message, preferring instead to have it presented in a manner that is non-offensive and nevertheless understandable in a polite tone.

On the other hand, many people might be persuaded to refrain from using profanity if they were given the option to publish either a less objectionable version of the message or a warning that it would be blocked altogether if it was uploaded.

4. Offensiveness Content in Social Media

Many online social networks use a variety of methods to prevent offensive posts from being published. When activated, Youtube's safety mode, for instance, prevents users from seeing any comments that include profanity. Clicking "Text Comments" will still show pre-screened text with offensive words replaced by asterisks. Facebook also allows users to create a "Moderation Blacklist" by entering keywords separated by commas. Use of blacklisted terms in a post or remark will result in the post or comment being flagged as spam and removed from the page. Apple Corporation did not approve of the "Tweetie 1.3" Twitter client because it allowed users to send tweets containing profanity. Twitter claims that users can simply ban and unfollow unpleasant posters if they see such posts, hence it currently does not pre-screen users' submitted contents. Most popular social media platforms rely on a basic lexicon-based method to filtering inappropriate information. Some, like YouTube, have predetermined dictionaries, while others rely on user contributions (such as Facebook).

In addition, the majority of sites rely on reports of inappropriate content from users before taking any action. These systems have limited accuracy and may produce numerous false positive alarms because they rely on a simple lexicon-based automatic filtering strategy to block the objectionable words and sentences. In addition, these systems frequently miss opportunities to act promptly when they rely on users and administrators to discover and report inappropriate content. Adolescents, who frequently lack cognitive understanding of risks, are particularly vulnerable to exposure, and these methods are unlikely to be helpful in protecting them from harm. In order to safeguard their children from being exposed to foul, pornographic, or hostile language, parents require more advanced software and detection methods.

5. Techniques to Detect Online Offensive Contents

Because the textual information in a social media context is typically unstructured, casual, and misspelled, identifying offensive language is a challenging endeavor. Researchers have researched smart techniques to identify offensive items using a text mining approach, as the existing defensive methods deployed by social media are insufficient. The following steps are necessary when using text mining methods to examine web-based data: There are three stages: 1) gathering data, 2) extracting features, and 3) classifying that data. The next sections will focus on the primary difficulties associated with employing text mining to detect offensive materials, which lie on the feature selection phrase.

a) Message-level Feature Extraction

Most offensive content detection research extracts two kinds of features: lexical and syntactic features.

Lexical features

In lexical analysis, each word or phrase is considered separately. Word frequency and keyword occurrence patterns are common ways to illustrate the language model. Bag-of-Words (BoW) was originally utilized for offences detection in earlier studies (McEnery et al., 2000). While analyzing a text, the BoW method simply counts the number of words without taking into account their context or meaning. Unfortunately, the BoW technique alone has limited accuracy in detecting subtle offensive language and also results in a significant false positive rate, especially during heated discussions, defensive responses to others' offensive remarks, and even talks between close friends. Since the N-gram method also takes into account the surrounding context of the words in order to identify potentially offensive material, it is seen as an improvement over previous methods (Pendar, 2007). N-grams are sequences of words inside longer texts that contain exactly N repetitions. Most text mining projects use N-grams of size two or three. N-gram, on the other hand, has trouble locating pairs of words that are closely linked yet widely spaced apart in texts. If N is increased, the issue is solved, but the system's processing performance is slowed and additional false positives are generated..

Syntactic features

Although lexical features perform well in detecting offensive entities, they are unable to differentiate the offensiveness of sentences that contain the same words but in different orders because they do not take into account the syntactical structure of the entire phrase. Consequently, natural language parsers (Marneffe et al., 2006) are

introduced to parse sentences on grammatical structures prior to feature selection in order to take syntactical features into account. Using a parser can aid with offensiveness identification by preventing the selection of irrelevant word sets as features.

6. Conclusion

The conclusion is that offensive language is a major source of trouble for online communities due to abusive behavior. Several machine learning models have been used in previous work to try and figure out how to spot abusive conduct. Users who intend to publish offensive content may be persuaded to rethink their decision if they are given the option of posting a more tame version of the same message alongside an alert that the offending content will be blocked.

References

Marneffe, M.-C. d. B. MacCartney, and C. D. (2006). Manning, "Generating typed dependency parses from phrase structure parses," in LREC.

McEnery, A. J. Baker, and Hardie, A. (2000). "Swearing and abuse in modern British English," in Practical Applications of Language Corpora Peter Lang, Hamburg, pp. 37-48

Pendar, N. (2007) "Toward spotting the pedophile telling victim from predator in text chats," in Proceedings of the First IEEE International Conference on Semantic Computing, pp. 235-241.

Xiang,G. ,Fan, B., Wang, L., Hong, J. and Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pages 1980–1984

Warner, W. and Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*. pp. 19–26.

Wang, W., Chen,L., Thirunarayan, K. and Amit, P.(2014) . Cursing in English on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. Pp. 415–425.

Nobata, C. Tetreault, J., Thomas,A., Mehdad, Y. and Chang, Y. (2016). Abusive language detection in online user content. *In Proceedings of the 25th International Conference on World Wide Web*. Pp. 145–153.

Burnap, A. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

Davidson, T., Warmsley, D., Macy, M. and Weber, I (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*. pp. 512–515.

Founta, A. M. D. Chatzakou, N. Kourtellis, J. Black- burn, A. Vakali, and I. Leontiadis. (2018). *A Unified Deep Learning Architecture for Abuse Detection*. ArXiv e-prints .

Wajnryb, R. (2005). *Expletive Deleted: A Good Look at Bad Languag*e. New York: Free Press.

Murray, J., Bradley, H., Craigie, W. and Onions, C. (1884). *Oxford English Dictionary*. Oxford: Oxford University Press.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. (2020). Offensive language and hate speech detection for Danish. In

Proceedings of the 12th Language Resources and Evaluation Conference (LREC).

Waseem, Z., T. Davidson, D. Warmsley, and I. Weber. (2017). *Understanding abuse: A typology of abusive language detection subtasks*. arXiv preprint ar Xiv:1705.09899.

Hamm, M. P., A. S. Newton, A. Chisholm, J. Shulhan, A. Milne, P. Sundar, H. Ennis, S. D. Scott, and L. Hartling. (2015). Prevalence and effect of cyberbullying on children and young people: A scoping re- view of social media studies. *JAMA*. 169(8):770–777

Kowalski, R. M. and S. P. Limber. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1):S13–S20.