ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

Detection of Curse Words and Hate Speech from Scrambled Words

Miftahul Islam[1] and Muhammad Mahfuz Hasan[2]

Abstract

Scrambled words that are formed due to the disorganization of one or more internal letters of a word have been analyzed by psycholinguists to be discernible by most of those who know of the language it is used in. These scrambled words can be generated by unmoderated users for spreading hate speech, curse words and expletives with the specific intent of bypassing existing censor filters. This paper presents a proposal to bolster detection of curse words and hate speech from internally scrambled words along with additional user preferred censor words. Our model has been developed to make general users safer from the atrocities of manipulated hate speech and curse word usage. Further improvements have been made to the detection of user manipulated offensive expletives through the use of permutations and their proper censorship and also through the application of different filters in various services used both in online and offline text-based media.

Keywords: Natural Language, Online Interaction, Swearing, Cursing, Social Media, Scrambled Words

[1]Student, Department of Computer Science and Engineering, Eastern University
[2] Associate Professor and Chairperson, Department of Computer Science and Engineering, Eastern University, Email: mhasan@easternuni.edu.bd

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

1. Introduction

The use of various mediums to communicate online; be it social media, online gaming, forum boards or messaging services have now vastly spread among us to communicate with each other. People are using these social networking sites because of their availability, flexibility and user-friendly features (Kumari & Verma, 2015). In this era of modern technology, text-based media acclaims as one of the widely used methods to interact with people through various Social Media Services, Online Forum Boards, Open-Closed Chats etc. (Cover, 2015). It has been reported that there are staggering amounts of hate speech, curse words and expletives generated by mostly unmoderated users, those are manipulated to bypass certain set filters for intelligent devices yet understandable by humans (Malmasi & Zampieri, 2018).

About 4 billion people are active in various forms of social media. Research also shows that in 2020, the average time spent on social media per day is 6 hours 42 minutes globally for users on any device (Sayimer, n.d.). People from different cultures, religions, genders, regions or even mentalities can be found on these platforms. Some of these people share their views and feelings without thinking about others, as flexibility and 'freedom of speech' is touted as one of the many points that let them do exactly that. Among them some contents may hurt someone's emotion or identity (Al-Hassan & Al-Dossari, 2019, February). Therefore, scrambled words exist due to user manipulation might not be discernible to anyone who does not have mastery over the language used; especially for those who do not have that same language as their mother tongue or even second language and so on. However, according to psycholinguists and their research, it is still possible for anyone to make these particular scrambled words legible just from a glance because of how they are scrambled to begin with. Although as of now there have been many ways to censor hate speech, usage of expletives and curse words, the same can also be said for methods of users bypassing the filters outright (Tsesis, 2001).

The use of various online communication tools in the form of mainstream social media, online gaming chats or even just anonymous forum boards has become vast in the last couple of years. Thus, people from much different psychological and diversified standing have come together with the ability to communicate with each other quite easily. This also brings about collisions between people of different mindsets. Since there is a sense of freedom of speech in these spaces, people also have the opportunity to express their views without thinking so much about ethics (Al-Hassan & Al-Dossari, 2019, February). This itself has propagated the spreading of hateful and offensive language. To be able to control and monitor the natural language that users apply can be an enormous task; especially with the rising number of users themselves. Even so there are now filters for hate speech, curse words and expletives in place. However, users have come

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

up with ingenious ways to bypass these filters in such a way that it cannot be detected by the already placed filters and therefore cannot be identified as offensive or hateful (Salloum *et el* 2017).

Considering the current situation this research intends to improve the identification of hate speech, curse words and expletives in various forms of text-based media so it can be made safer and more secure from any sort of hateful or offensive language in terms of their content. Main goal of this research work is to propose a set of data permutations of already existing censored words into the existing list of censored words used by various media and even on an individual basis (Twitter and Facebook let users block out usage of user-set words) (Salloum *et el* 2017; Burnap & Williams, 201). This updated list of words will find user manipulated language that has been scrambled enough to bypass the already placed filters for censored content yet be legible to the person who reads it.

It has already seen use in the modern internet era, especially in social media, online forums and online gaming text chat communications. Social media is as powerful as of late that anything can be 'viral' in a moment's notice. It has already been acknowledged that hate speech can incite in the real world pertaining to race, nationality, gender, religion etc. and its subsequent possible horrible outcome like riots, general unrest, terrorism; even suicidal tendencies on an individual level. These situations can be deterred by nipping them in the bud which sometimes forms from various calls to action online through hate speech. It will be a huge undertaking for anyone involved, considering the number of ever-growing users and their numerous content; even the complexity of natural language and its many new ways of using it is at play. It is the hope of these researchers that even with this small step, further betterment of censoring hateful and offensive language in text-based media will be made to have a stride now and in the future.

This research provides a psycholinguistics-based approach that shows the possibility of users being able to understand other user-generated offensive and hateful language that has been scrambled in such a specific way that it bypasses the already placed text censor filters used by various media. Moreover, for the small number of text-based content that bypass the systems in place, we can further use our proposed method to pinpoint them even with the use of scrambled internal letters of the words in the text-based content.

2. Curse Words and Hate Speech Detection in Existing Systems

This section explores the status of online hate speech and its social control from institutional and corporate perspectives. Hate speech has increasingly become a source of societal and political concern across the globe, as witnessed by recent measures and initiatives to tackle it. Hate speech is clearly a global phenomenon and one which rests on some form of inter-group hostility. The following subsections describe some of the social media, which have their own policies to prevent the speeding of hate speech.

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

2.1. Facebook

Facebook has its policy about defining what can be considered as hate speech. Any content that directly attacks people based on various protected characteristics such as race, ethnicity, nationality, religion, gender, disability and/or disease can fall under hate speech. Direct attacks also can include dehumanizing speech, harmful dismissal, stereotype profiling, calls for seclusion and/or segregation etc. These are all well explained in Facebook's Transparency Center. However, Facebook rules do not forbid the use of curse words and/or swearing when not used for hate speech. These can be used in posts, comments and messages. To this end, manual moderation for certain pages and user profiles is required to take out any profanity that is used in user-generated content shown. In addition to this, certain business-related pages already have the option to censor any word that they might like through the tool provided by Facebook for the administrator of that page (Ben-David & Fernández, 2016).

2.2. Twitter

The user policy for Twitter states that username, display name, or profile bio must not be used to engage in abusive behavior such as targeted harassment or expressing hate towards a person, group, or protected category. This mainly applies to content that incites fear, reduction of someone to less than a person, violent threats, abusive slurs, hateful epithets, racism and/or sexist tropes. In terms of using strong language like expletives and curses, Twitter now has updated its stance to take into account who you are interacting with your content. This update now takes into account the relationship users have with the people they interact with and therefore can use general cursing in said text-based content; mainly messaging. Twitter also has a partnership with Cardiff University's School of Social Sciences and School of Computer Science and Informatics. This partnership is a part of the University's School of Social Data Science Lab and is called "HateLab". They developed the HateLab Dashboard to use internally which uses the Twitter API and machine learning to classify toxic speech (Burnap & Williams, 2015; Alkiviadou, 2019).

2.3. YouTube

According to YouTube policy, any form of hate speech is not allowed on the platform. YouTube has and will remove content that condones and/or promotes hate, violence or demeaning based on age, caste, ethnicity, race, religion, gender, disability etc. In terms of use of cursing and profanity, this is only relegated to people in the YouTube Partner Program and their advertiser-friendly guidelines. Even then there is a very loose leash for the content creators to have usage of profanity in their content according to some reports. These reports also suggest that there are selective cases for the usage of profanity and it is at various financial advertisers' behest to have the relative content monetized or not; but as long as the content does not incite any sort of hate speech, YouTube will host the content (Alkiviadou, 2019).

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

2.4. Reddit

Among the existing online forums most of them on the internet and are still relatively popular in terms of their somewhat mainstream audience, Reddit always comes out on top. Home to thousands upon thousands of communities on nearly anything and everything, the site has nearly more than fifty billion views on a monthly basis with more than 52 million active users daily. With such a broad audience, there are many policies in place for ensuring security and anonymity. However, with that anonymity come a lot of other precedent problems as well, from the use and abuse of unlawful content to even condoning and propagating hate. Reddit was and still is touted as one of the most difficult-to-regulate platforms on the internet, with a low reputation as being a "cesspool of racism". Back in June 2020, Reddit updated its policy on hate speech (Rieger *et al.*, 2021). The enforcement of the policy led to almost two thousand "subreddit" communities being banned from the platform outright. Online forums like these on the other hand do not have any policy on swearing and use of profanity on their platform unless specified by moderators of certain forums; in Reddit's case, subreddits. Recently in 2017, a research group in Germany had gone through thirty-five different subreddits that were most popular in seven different categories and found that across over thirty million comments; almost 11.93% of them had at least one use of profanity (Thomä, 2017). This is somewhat skewed due to the fact that it takes into account the "Karma" score that Reddit gives access to its users where they can essentially like or dislike a post or comment. Nonetheless, this goes to show that even though users do always have the intent of using hate speech, they do have the tendency to swear in terms of normal posting of comments on Reddit at least 10% of the time.

2.5. Various Online Gaming Media and Chat Services

Online gaming has become very popular in the last couple of years and so has its reputation. However, due to some ill-minded individuals, many users in this space can be adversely affected mentally and even in some cases, physically. In terms of policy, most of these online games prohibit the users from using any and all forms of hate speech (Sublette & Mullan, 2012). Some online games in accordance with their age group and rating also prohibit them from using any sort of profanities in the communication end of the game. It has been reported that online gamers have some sort of harassment in their gaming sessions and it was not only in their online sessions but in their in-person LAN parties or gaming clubs as well (Saarinen, 2017). In most online games now, there are options enabled for the users by the developers to either mute, block and/or black list another user that is being hateful or harassing towards them. There are options to report such abusive users to the developers as well where it can be manually reviewed, resulting in the perpetrator receiving a temporary or permanent ban from those online games(Saleem *et al.*, 2017).

In terms of various online chat services that are accessible using browsers or as apps on mobile devices, there are some strict policies for hate speech in public chat rooms. This, however, is lax

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

in private chat rooms and left to the moderating users' discrepancy; profanity also adheres to the same rules in this case. In terms of chat services for various services like a live chat for customer service; if done in a text-based format, it is obvious that hate speech is not tolerable in any format. Of course, it must be said that adolescents should not use profanity at all but in terms of people of age, some see it as a freedom of expression; however, it should not be touted amongst everyone and should possibly be kept among close peers and in private.

3. Proposed Algorithm

Our main proposal is that permutations of any censorable words based on their scrambled internal letters be used in the filters made to detect hate speech, curse words or expletives to further strengthen the already placed systems for detecting and censoring hateful and offensive text content that various social media platforms, online gaming text communications systems and public chat spaces use. This can even be done on the entire individual censor word, but that is left to the discretion of the developers and if permitted by them, the users as well. We can have developers of various the many aforementioned text-based services integrate this method of permutation for their already existing systems seamlessly.

Therefore, our proposal may be summed up as the following:

- The proposed algorithm may be added to the existing systems for censoring unwanted text-based content.
- Let users add their own personal words that they find problems with to the list of censor words for added security against any personal targeting using slurs, slang, ethnicity etc.
- Have the newly censored content be analyzed for further scrutiny into finding out how the bad actors are playing their role in using text-based content

In some cases, this will be done differently as there are services that are already developed across a multitude of programming languages which currently are active but the core concept remains the same. This is mainly done for the text-based user-generated content that is most of the time unmoderated due to the sheer number of them existing across any single individual platform service. Though there is automated moderation which can take care of most of the heavy lifting in terms of censoring the targeted hate speech and offensive language, we have already seen how it can be bypassed through the use of scrambled words.

3.1. Pseudocode

The proposed method of this research work can be broken down into the following steps:

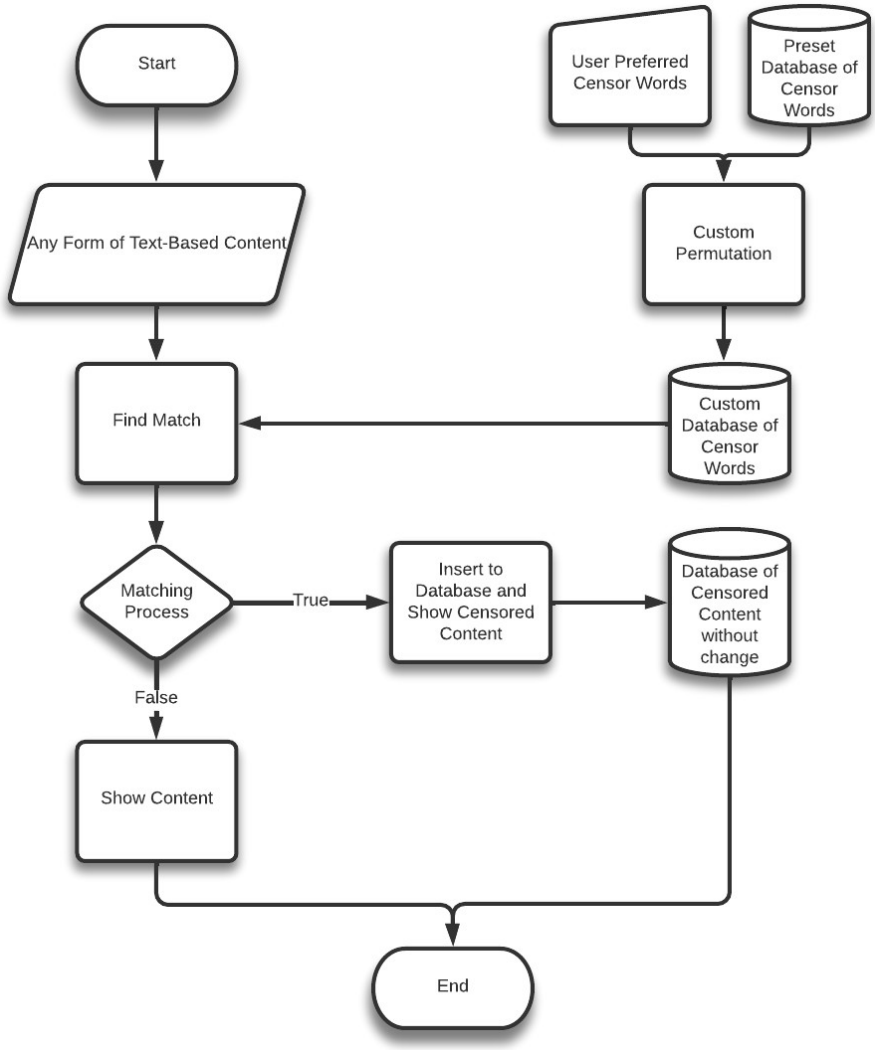
- The process starts with the notion that there will be user-input text-based content that will be put through to be matched for censored content from the database.
- Generate a custom database that is comprised of the set custom permutations of the user-

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

preferred censor words and existing preset database of censor words; the permutation of the censored words will be based on that of the internal letters and keeping the first and last letter of the word static.

- Take the user input and match it against the custom database.
- If the process finds a match, store the matched content without modification into a new database for reference and further analysis, censor the content that has been matched and show the censored content
- If no match, the content is shown without any modification.
- The process ends.



ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

Fig.1: Flow Chart for Censor Word Recognition

4. Findings from Survey

An online survey has been conducted among 253 participants using Google Forms. The outcomes of the responses to the survey are as follows:

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

Question 1: In your everyday interactions online (Facebook/Twitter/Youtube/Reddit etc) and offline (conversations/reading/writing/listening etc), HOW MUCH DO YOU ENCOUNTER Hate Speech, Curse Words, Expletives (both mild and/or severe)?

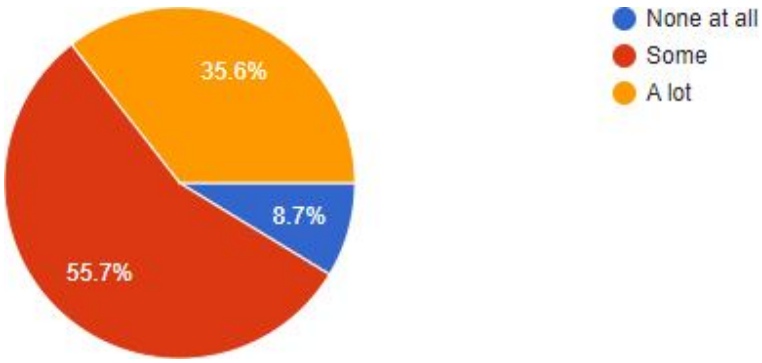


Fig.2: First Survey Question and Answer Pie Chart

Question 2: In your everyday interactions online (Facebook/Twitter/Youtube/Reddit etc) and offline (conversations/reading/writing/listening etc), HOW MUCH DO YOU USE Hate Speech, Curse Words, and Expletives (both mild and/or severe)?

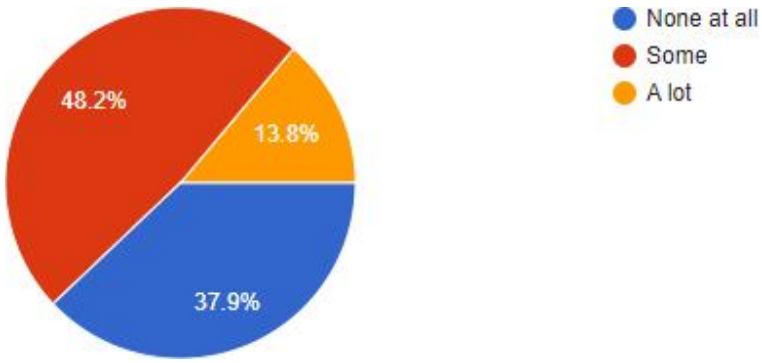


Fig. 3: Second Survey Question and Answer Pie Chart

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

Question 3: DON'T WAIT AND DO INSTANTLY, read the following: The human mind does not read every letter by itself, but the word as a word

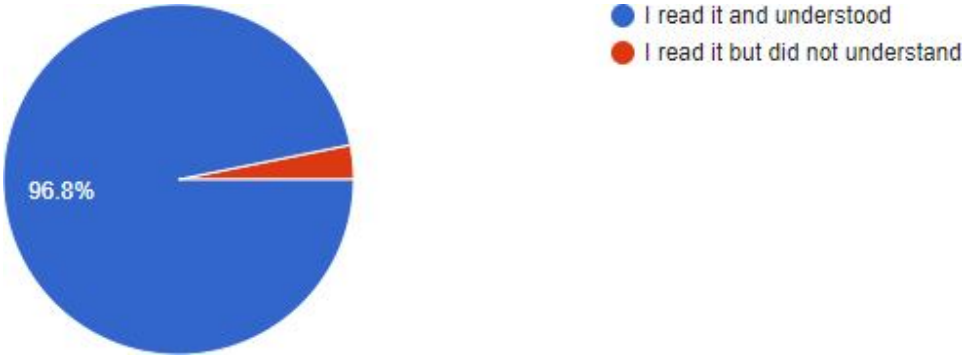


Fig. 4: Third and final Survey Question and Answer Pie Chart

In the survey, participants were prolific users of the English language and they were asked three separate questions from which the researchers tried to understand how much do the people themselves encounter and/or use profanity and/or hate speech.

These questions were asked to understand the frequency of people getting thrown profanity at them or if they themselves were the ones throwing it. Of the participants, only 8.7% of them claimed to have never encountered any hate speech or expletives at all and 37.9% of them claimed to have never used any themselves.

However, the main reason for this survey was to ascertain if they could understand simple internally scrambled words and surprisingly a massive 96.8% of the participants were able to clearly grasp the meaning without spending any time. This provides an indubitable weight towards our dissertation.

5. Conclusion

As netizens of the internet, we use English as the primary language to overcome communication barriers across people who speak different languages. With its incredible usefulness come many caveats too. One of them is the fact that the same kinds of words can have different meanings in different languages. We are mainly focusing on the English language for this dissertation but other languages can be used to have localized effects nonetheless. As an example, on many social media platforms, comments are made to have some expletives that are written in scrambled forms or with phonetic letter replacements and more.

This research work emphasized the process of better identifying and finding a solution to the

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

abuse of certain scrambled words that bypass the various censor filters for offensive, derogatory and hateful language. By nature, these scrambled words can be understood with our human psychological functions, achieving the objective of the perpetrator who used them in those scrambled forms to incite their vile goals. To this end, we have proposed the use of permuted words from the existing censor filter database with the addition of custom user-set censor word filters that some platforms already use and their permutations. The internet, online platforms and people having conversations in a shroud of complete anonymity are still young in the grand scheme of human civilization and this in itself can mean that new conversational or language-based habits both good or bad could develop the Inclusion of this proposition most certainly can and will reduce some vitriol in online spaces in terms of text-based communications.

6. Future Works

In terms of language, English is predominantly used as the premier language to go through most language barriers. As of now, there exists very few methods of understanding hate speech and offensive language if someone is using a different language that other users or other development companies don't understand. One example can be given from back in 2018 when Sri Lanka had to block all access to Facebook, any form of social media and messaging applications. At the time, the Sinhala language was used to incite an anti-Muslim campaign and it could not be understood by Facebook to be marked down as hate speech.

Any and all use of hate speech is always a bad thing, this much is true and has always been; however, the use of profanity can towards other known acquaintances. This is why it is very much dependent on the context of using the profanity. To this end, artificial intelligence-based keyword filtering with permutations in text editors and other platforms can be a way to alleviate some of the work. Research even has shown that swearing can increase the-pain tolerance in people. There are even some forms of profanity that may have been considered taboo from a religious standpoint in the past but now is a common term used in everyday conversations; an example being "n Nmy god!".

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

References

Al-Hassan, A., & Al-Dossari, H. (2019, February). Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology* (VBen-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 27.

Alkiviadou, N. (2019). Hate speech on social media networks: towards a regulatory framework? Information & Communications Technology Law, 28(1), 19-35.

Burnap, P., & 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.

Cover, R. (2015). Digital identities: Creating and communicating the online self. Academic Press.

Kumari, A., & Verma, J. (2015). Impact of social networking sites on social interaction-a study of college students. *Journal of Humanities and Social Sciences*, *4*(2), 55-62.

Malmasi, S., &Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, *30*(2), 187-202.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media+ Society*, *7*(4), 20563051211052906.

Saarinen, T. (2017). Toxic behavior in online games. *Unpublished master's thesis, The University of Oulu, Oulu, Finland*.

Saleem, H. M., Dillon, K. P., Benesch, S., &Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Salloum, S. A., Al-Emran, M., Monem, A. A., &Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

ISSN: 1999-7889
Volume 10, Issue 1 & 2 (December 2022)
March 2024

Eastern University
The Eastern University Journal
A UGC Approved Peer Reviewed National Journal
www.easternuni.edu.bd

Sayimer, I. Social media effect on corporate communications: A survey study of young adults' social media habits in turkey.

Sublette, V. A., & Mullan, B. (2012). Consequences of play: A systematic review of the effects of online gaming. *International Journal of Mental Health and Addiction*, *10*(1), 3-23.

Thomä, J. (2017). *Swearing in a public place: on the usage of swear words on reddit* (Doctoral dissertation, Universität Potsdam).

Tsesis, A. (2001). Hate in cyberspace: Regulating hate speech on the Internet. *San Diego L. Rev.*, *38*, 817.