Journal of Engineering Science and Technology
Special Issue on ICITE2021, October (2021) 37-46

A MACHINE LEARNING MODEL FOR THE PROFANITY DETECTION IN THE FILIPINO LANGUAGE

BEVERLY P. FERRER*, CRYSTELLE T. TOMILAS, LORENZ P. MALLARE, BENEDICT D. PINEDA. JR., ANA F. DE GUZMAN, FAMELA N. SIAPNO, KATHLEEN B. PAYANG, MICHAEL B. YU, ALLEN D. BOLAÑOS, ZAREENA D. LEE

School of Accountancy, Management, and Information Studies, Saint Louis University, Baguio City, Philippines
*Corresponding Author: bpferrer@slu.edu.ph

Abstract

In online environments where profanities are not properly regulated, many users cannot avoid exposure to the content found therein. Due to the freedom of speech that social media entails, some users make use of foul language. To prevent this, applications have been developed that can filter out profanities, but none have targeted the Filipino language, which is crucial given that many Filipinos are social media users. For this research, the team was able to successfully build a machine-learned model that can classify whether a text contained Filipino profanities or not. After performing descriptive analysis, it was determined that Filipinos tend to use profanities when exhibiting strong expressions that are not necessarily negative, such as feelings of admiration or yearning. However, some users also expressed profanities negatively towards matters that they find hard or irritating. Lastly, a prototype web application was also developed which integrated the predictive model. To obtain a more diverse set of topics for the descriptive analysis, the team suggests scraping tweets within a larger time frame. As for the predictive model, the team acknowledges that it can be further improved by integrating Semantic and Sentiment Analysis to more accurately target social media posts that deliberately berate others.

Keywords: Filipino, Latent Dirichlet Allocation, Machine-learning, Profanity, Text classification.

1. Introduction

Profanity has, in a way, become a major part of language and speech due to its aid in expression. It has been a prevalent issue in society since the early 1900s. Unfortunately, profanity can harm a person's thoughts. Certain words in a language have been considered profane by certain communities, and the use of such a type of language has been discouraged by others. Harsher language may create negative associations because it is both unexpected and unpleasant to the receiver [1].

In an online setting, particularly, profanity can often be encountered during gaming sessions. A study by Pujante [2] suggests that the "trash talk" (i.e. swear language) of gamers is more of a way for them to show off their wit and creativity rather than for insulting others. However, even though it is perceived to be normal for gamers to use profanity when gaming, negative reactions toward these words could arise especially for those who are not used to such language [3]. Profanity is also often used in cyberbullying, a violent and abusive activity that is done by sending or posting offensive content. Most cyberbullying victims are teenagers and 36.5% of students had encountered instances of cyberbullying at least once [4].

A 2014 Pew Research Center study found that about 22% of Internet users had been victims of online harassment in the comment section of a website [5]. This problem can be very detrimental to a country with a large internet market. Internet users in the Philippines increased by 4.2 million users between 2020 and 2021, and as of January 2021, there are 73.91 million internet users [6]. Such a high user count would account for many instances of profanity, a common method to give negative comments and feedback to others. In the context of social situations, profanities include swear words that are considered improper or unacceptable, particularly characterized by obscenities, blasphemy, or offensive slurs [7].

The Internet has been a propagating place for new forms of mental and emotional disturbance, which is why appropriate measures should be in place to minimize these. Numerous studies have already been made on the topic of abusive language detection on social media sites and other online platforms; however, the majority of such only focused on detecting keywords with some content analysis based on a selected collection of existing texts or datasets [8]. In the following paragraphs, we will discuss some models that have been created to detect profanities or abusive language in general in an online setting.

According to Modha et al. [9], social media sites alone contain massive amounts of hate speech conversations/comments and offensive posts which aim to insult or degrade other individuals or groups on the platform. Thus, such sites implement their content moderation system to detect such foul language or interactions; however, such tools are not readily available for use by regular online users and law enforcement agencies who monitor online activities. Thus, in their study, they had designed their own hate speech detection tool, which serves as a browser plugin that would visualize aggressive or offensive content on social media sites. Their study made use of the Support Vector Machine and the Convolution Neural Network.

Alakrot et al. [10] conducted predictive modeling to accurately detect offensive behavior in online communication. They gathered data from comments on sites like YouTube. However, the data gathered was only focused on the videos that are top viewed or controversial since according to the researchers, these are the types of videos that often invoke toxic and offensive comments from the viewers.

Additionally, the dataset that was used was created in the context of their local language, which is Arabic.

As years passed by, machine learning approaches shifted to deep learning approaches. Most of the machine-learned models are targeting social media and online games where bad words are very common in chat environments. A study by Alshalan and Al-Khalifa [11] tackles the deep learning approach for automatic hate speech detection. Their study constructed a public dataset of 9316 tweets where each tweet consists of abusive words. However, their study was only targeted toward Arabic tweets.

The observed problem that motivated the topic for this study is the increase in profanity in day-to-day conversations of people on the Internet and the impacts of profanity that can affect our everyday lives. It is much easier now for everyone to express these profanities through social media and messaging platforms. Profanity can be interpreted as unprofessional, disrespectful, offensive, or even harmful when used in the wrong environment. Communicating with the use of profanity could harm others [1].

Knowing that there is a massive number of Internet users in the Philippines, the growth of profanity is unavoidable. Unfortunately, there is a lack of available studies that focused on creating profanity filters, particularly for the Filipino language. Thus, many online environments are still not able to censor those profanities. Profanity in the Filipino language needs to be addressed as it can be deemed offensive or inappropriate, especially to those who understand it. There are some existing applications for detecting profanity in chat environments but they are only limited to detecting other languages such as English, Hindi, and Arabic.

The main goal of this study, therefore, is to create a prototype of a web application that is intended for educational purposes and automatically detects posts that contain profanities. Moreover, the team also aims to perform a descriptive analysis of the extracted profane Filipino tweets with the use of Latent Dirichlet Allocation (LDA) to analyze possible topics related to the use of profanities on Twitter. The specific objectives of this project are to:

- Develop a machine-learned model to detect profanities or abusive language in the Filipino language.
- Describe the topics related to the use of profanities.
- Integrate the developed machine-learned model in a web application that filters and prevents posts with profanities.

Since the prototype web application will be of an educational type, it is expected that users maintain professionalism and avoid the use of foul language. With the use of the model to be developed, posts that are detected to contain profanities will not be posted, which would ease the job of post moderators since they would not have to manually review each post to check whether it contains profanities. Moreover, this model can be integrated with other web applications, especially since there is a lack of data models that can detect Filipino profanities. With this, the web applications themselves will be able to moderate user posts, comment sections, and chat environments. This can also help members of online communities by lessening the toxicity that is caused by profanities. Furthermore, the results of the descriptive analysis can also provide some insights as to what topics or words people usually use along with profanities.

2. Methodology

The following sections describe the methods to achieve the objectives of the study.

2.1. Developing the machine learned model to detect profanities and abusive words in the Filipino language

Initially, for the data-gathering phase, the team had searched for possible word lists that contained Filipino profanities which would serve as the main basis for determining if certain tweets or text contain such profanities. After evaluating the list, the team then used the Filipino-badwords-list created by Estiller et al. [12] which was then incorporated to formulate the CSV file containing a lexicon of common Filipino profanities.

For the process of data collection, the team decided to use Twitter as the main social media site for the source of data. Using TWINT, the team scraped tweets in real-time and filtered them based on the aforementioned list of profane words in Filipino. Tweets not containing any Filipino profanities were also extracted which would be used later on in the labeling process. The time frame from which the tweets were extracted is from 1:40 PM to 12:40 AM on May 23, 2021. These tweets were then collated into a CSV file that served as the main dataset of the model consisting of an estimated 11,483 tweets.

Data preprocessing was done after collecting the dataset of tweets, where the team cleaned and prepared the data for further processing. During the preprocessing stage, all alphabetical characters for each tweet were converted to lowercase. User tags (i.e. usernames of users tagged in a tweet using the '@' symbol) and URLs (Uniform Resource Locator) were also removed. Through the use of the Python langdetect library, the team was able to filter out tweets that are only in the Tagalog language which remained in the final dataset.

The process for data labeling was automated through the use of Python codes and a list of Filipino profanities. The team programmed the automation to label a tweet as "negative" if any Filipino profanity from the list was present in the tweet; otherwise, it would be labeled as "positive". The team then reviewed the resulting dataset to ensure that it was labeled according to what was intended.

The dataset was then balanced using a worksheet application by deleting some rows at random until there was an equal number of rows for tweets under the negative and positive categories. The dataset contained a total of 8,734 tweets after balancing the dataset with 4,367 "negative" tweets and 4,367 "positive" tweets, which were used for the training of the model. This is then the cleaned, labeled, and balanced dataset that will be used for the machine-learned model as well as the descriptive analysis.

The project implemented supervised learning in the process of performing text classification techniques. In training the machine-learned model, for the possible classification models that were compared, the team made use of the following classification techniques: Naive Bayes, Random Forest, Gradient Boosting, and Support Vector Machine.

The dataset of tweets was then split into the training and testing datasets (80% for training and 20% for testing) [13]. The classifiers were trained and tested separately. After the training and testing of each classifier, a classification report

was outputted to determine the respective performance metrics of the models. The team then compared and evaluated the performance metrics of each of the classifiers and then selected the most appropriate model with the best performance metrics (Accuracy, AUC-ROC, Cross-Validation. Moreover, the Confusion Matrix was used to further analyze the results of the model. After selecting the classifier with the best performance metrics, a machine-learned model was developed utilizing the selected classifier model.

2.2. Describing the topics related to the use of profanities

For the process of descriptive analytics, the project implemented unsupervised learning which involved topic modeling. The dataset used was taken from the cleaned, labeled, and balanced dataset. Firstly, the dataset of negative tweets was extracted and was used as the dataset for the analysis. Positive tweets were not included since the concern for the descriptive analysis is to inspect common words that are in tweets containing profanities. English and Filipino stop words were then removed from the set of negative tweets. A list of Filipino stop words [14] was used to filter out texts that are unnecessary to the application.

The team then used GridSearchCV on the dataset to determine the best values for both the learning decay and the number of topics. The resulting values were then used as the inputs of the learning decay and number of topics for the Latent Dirichlet Allocation (LDA) model. The LDA model was used to determine the topics of the dataset as well as the top terminologies per topic. The results from the LDA model were then visualized using pyLDAvis, which shows the intertopic distance map and the most common words for each topic. Finally, each topic from the LDA results was analyzed by checking how the most common words in that topic were used in the dataset of negative tweets.

2.3. Developing the web application prototype for a student discussion board

Finally, the team created a prototype website using Anvil, which integrated the developed machine-learned model. The prototype is a web application that automatically detects and filters posts that contain profanities on a student discussion board.

The prototype website was uplinked to the Jupyter notebook that contains the codes for the model. This was done by installing the Anvil Uplink library in Jupyter, importing the Anvil server, and then connecting the notebook by adding the Anvil Uplink key to the codes in the notebook. Accordingly, the model was integrated into the post-publishing feature of the website wherein if the post contains any Filipino profanities, it would not be published and the user would be notified; while posts that do not contain profanities will be published in the main feed of the forum.

3. Results and Discussion

3.1. Developed machine learned model

The accuracy score, ROC-AUC score, cross-validation score, and the confusion matrix were used to determine which classifier will perform the best in determining whether a tweet contains Filipino profanities or not. A higher score would mean that the classifier performed better. The results of classifying the data set for the training of the model are displayed in Table 1. The dataset was split and 80% was allocated for the training data while 20% was allocated for the testing data. The

table provides the different metrics scores of the chosen models which vary from 0.5003 - 0.9954 in the accuracy scores, 0.5892 - 0.9995 in the AUC-ROC scores, and 0.5070 - 0.6098 in the 5-fold cross-validation scores. Five-fold cross-validation was used instead of the 10-fold cross-validation due to restraints in the computing power of the team's devices.

Table 1. Performance metric comparison of the four classifiers.

Classifier	Accuracy	AUC-ROC	Cross-Validation
Naive Bayes	0.5003	0.5892	0.5543
Random Forest	0.9760	0.9964	0.6098
Support Vector Machine	0.9805	0.9985	0.6098
Gradient Boosting Machine	0.9954	0.9995	0.5070

As for the confusion matrix, true positives and true negatives would have to yield a higher count while false positives and false negatives would have to yield a lower count. The resulting confusion matrix for each classifier is shown in Table 2.

Table 2. Confusion matrices of the four classifiers.

Classifier	Confusion Matrix			
Naive Bayes			PREDICTED	
			NEG	POS
	ACTUAL	NEG	866	0
		POS	873	8
Random Forest			PREDICTED	
			NEG	POS
	ACTUAL	NEG	853	13
		POS	29	852
Support Vector Machine			PREDICTED	
	ACTUAL		NEG	POS
		NEG	856	10
		POS	24	857
Gradient Boosting Machine			PREDICTED	
			NEG	POS
	ACTUAL	NEG	857	7
		POS	0	883

The team has analyzed that, among the four models that have been trained, the model that utilized the Gradient Boosting Classifier has the best performance because of its high accuracy score and ROC-AUC score of at least 0.99 and its confusion matrix that displays the highest number of true positives and true negatives and the least number of false positives and false negatives. Thus, Gradient Boosting Classifier was chosen to be integrated into the prototype web application to have a better chance of detecting Filipino profanity words.

3.2. Topics related to the use of profanities

Using GridSearchCV, the best values for the number of topics and the learning decay were determined. The resulting optimal value for the n_components is 5 while the optimal learning_decay value is 0.9. These were the values used for building the LDA model, which was then visualized using pyLDAvis.

Figure 1 shows the intertopic distance map generated by pyLDAvis. Circles that are larger mean that the topic has a higher percentage of tweet count in the corpus. The numbers are sorted from smallest to large, with Topic 1 having the largest percentage and Topic 10 having the smallest percentage. There is an overlap between Topics 1 and 3, which means that they have similar terms.

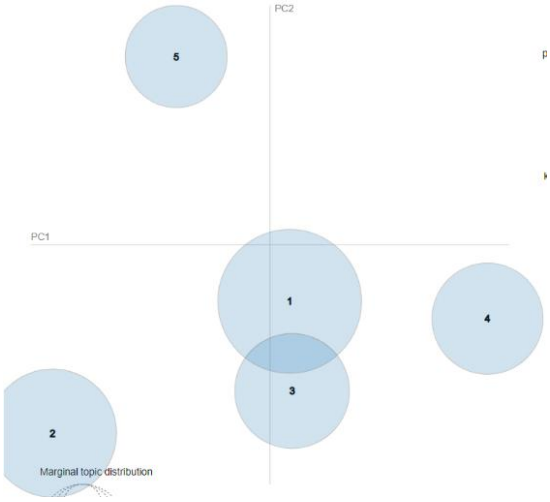

Fig. 1. Intertopic distance map.

Topic 1 is mainly about tweets involving profanities to express headaches. Notable terms here include the words "sakit" (pain), "ulo" (head), and "init" (heat). Going back to the dataset of negative tweets, the team noticed that users tend to use such terms on matters that they do not understand or on matters that irritate them. Among those that irritate them is the heat or "init" in Filipino, which has been a recent problem in the Philippines this May 2021 [15]. The terms "ganda" and "boses" are also in the chart, although there was very little relation of these terms to "sakit", "ulo", and "init". Moreover, the intertopic distance map shows an overlap between topics 1 and 3. After studying the tweets, the team observed that the terms "ganda" and "boses" were in the tweets having the term "sb" (to be explained in Topic 3).

Topic 2 is mainly about tweets involving profanities to express the feeling of missing someone. Notable terms here include the words "miss" (miss) and "kita" (you). The team noticed that users tend to use profanities to further emphasize the feeling of missing someone or something. Some tweets emphasized how users missed moments before the COVID-19 pandemic.

Next is topic 3, with terms such as "mahal" (love) and "kita" (you) in tweets where users tend to use profanities to further emphasize their feelings of admiration towards someone or something. Among these is a Filipino boy-band called SB19, thus the term "sb", "official" and "mapaonasap". During the scraping of tweets, SB19 recently

performed their song “Mapa” on ASAP, a Filipino show, and users were expressing their love for the band and the song. As mentioned in Topic 1, the terms “ganda” (beautiful) and “boses” (voice) occurred in the same tweets referring to SB19.

Just like Topic 3, Topic 4 mainly consists of tweets with profanities towards a certain someone. The notable terms here are “kinikilig” (feeling of romantic excitement), “cute”, and “naiiyak” (about to cry), which were used with the term “kyungsoo”, a member of a South Korean Boy Group named EXO. When the team checked the tweets, most of the tweets about Kyungsoo were positive even though profanities were used within the tweet.

In Topic 5, most of the tweets associated with the terms in the chart contain profanities toward a mobile video game. The term “cod” is present in this topic which is a shortcut for Call of Duty, a popular mobile video game. A notable term in this topic is “hirap” (hard) which indicates the users’ way of expressing their experiences when playing the game.

3.3. Web application prototype for a student discussion board

The web application prototype was created using Anvil to show a partial implementation of the web application. The process starts with inputting a username, followed by reading community guidelines, selecting among a list of discussion boards, and publishing a post on the selected discussion board. The user can publish a post on a selected discussion board (see Fig. 2). The post is classified by the model that is integrated into the web application. Posts that are classified by the model as “positive” are successfully posted on the discussion board, while posts that are classified by the model as “negative” are not posted and the user is notified about it.

Figure 2 shows a screenshot of the input box used to create posts in a discussion board of the web app prototype. When a user creates a post and it has been published successfully, a pop-up will appear, notifying the user that the post has been shared. However, if a user attempts to create a post with a blank title, blank content, or both, the respective pop-ups will appear. Moreover, a post can also fail if it goes against the community guidelines (i.e., contains profanity). This will cause the pop-ups to show if the post title or content contains profanity (see Fig. 3).

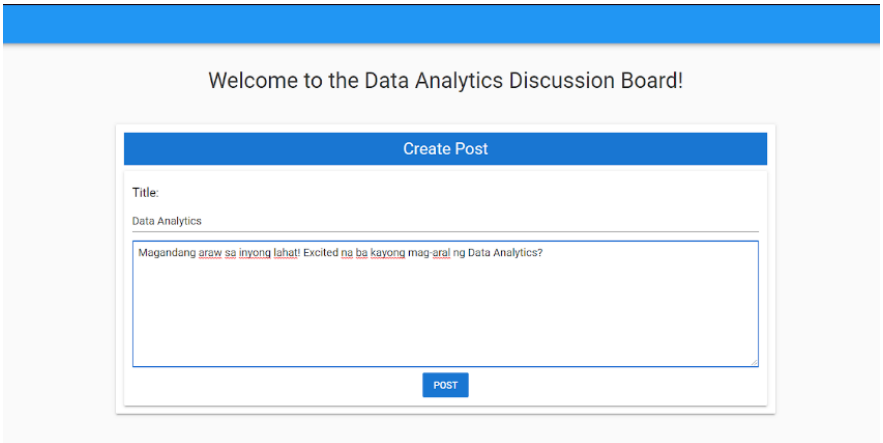

Fig. 2. “Create Post” interface.

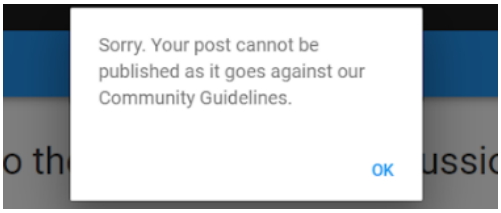

Fig. 3. Failed post popup (inappropriate title and/or content).

4. Conclusions

A machine-learned model that automatically detects profanities and abusive words in the Filipino language was developed. Among the classification algorithms used in developing the model, the Gradient Boosting Classifier yielded the highest performance metrics. It had an accuracy of 99.54% and an AUC-ROC score of 0.9995.

For the descriptive analysis, the team used the set of negative tweets and removed stop words. LDA was used to group related words into 5 different topics and was visualized using pyLDAvis. The team has analyzed that profanity can be used to express, not only negative feelings but positive feelings as well. Among these negative feelings were matters that they do not understand or get irritated by. Profanities were also expressed to emphasize their feelings of admiration towards particular artists.

With the use of Anvil, the team was able to create a prototype of a web application for a student discussion board. It integrates the machine-learned model that was developed to detect Filipino profanities. However, the prototype is currently limited to classifying whether a text contains Filipino profanity or not.

The team acknowledges that the created predictive classifier model can still be improved. For future studies, the model can be augmented by implementing a more in-depth classification feature, wherein the profanities can be categorized by severity or by purpose. However, this process would require validation from a language expert. Sentiment and semantic analysis could also be integrated wherein a statement that supposedly contains profanities could be analyzed. Moreover, a machine-learned model that can detect posts that include a combination of English and Filipino profanities can be developed. For better results on the descriptive analysis, the team also recommends scraping tweets from a larger time frame to include a bigger diversity of topics containing profanities. Lastly, the team recommends that the model should also be integrated for other applications such as social media and communication platforms.

Abbreviations

LDA	Latent Dirichlet Allocation
ROC- AUC	Area Under the Curve-Receiver Operator Characteristic

References

1. DeFrank, M.; and Kahlbaugh, P. (2018). Language choice matters: when profanity affects how people are judged. *Journal of Language and Social Psychology*, 38(1), 126-141.

2. Pujante, N.T.Jr. (2021). Speech for fun, fury, and freedom: exploring trash talk in gaming stations. *Asian Journal of Language, Literature and Culture Studies*, 4(1), 1-11.
3. Berowa, A.C.; Ella, J.R.; and Lucas, R.G. (2019). Perceived offensiveness of swear words across genders. *The Asian EFL Journal*, 25(52), 163-187.
4. Perera, A.; and Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, 181, 605-611.
5. Anderson, M. (2014). About 1 in 5 victims of online harassment say it happened in the comments section. Retrieved May 23, 2021, from https://www.pewresearch.org/fact-tank/2014/11/20/about-1-in-5-victims-of-online-harassment-say-it-happened-in-the-comments-section/
6. Kemp, S. (2021). Digital in the Philippines: all the statistics you need in 2021. Retrieved May 23, 2021, from https://datareportal.com/reports/digital-2021-philippines
7. Feldman, G.; Lian, H.; Kosinski, M.; and Stillwell, D. (2017). Frankly, we do give a damn: the relationship between profanity and honesty. *Social Psychological and Personality Science*, 8(7), 816-826.
8. Abderrouaf, C.; and Oussalah, M. (2019). On online hate speech detection. effects of negated data construction. *2019 IEEE International Conference on Big Data*, 5595-5602.
9. Modha, S.; Majumder, P.; Mandl, T.; and Mandalia, C. (2020). Detecting and visualizing hate speech in social media: a cyber watchdog for surveillance. *Expert Systems with Applications*, 161, 113725.
10. Alakrot, A.; Murray, L.; and Nikolov, N.S. (2018). Towards accurate detection of offensive language in online communication in Arabic. *Procedia Computer Science*, 142, 315-320.
11. Alshalan, R.; and Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the Saudi twittersphere. *Applied Sciences*, 10(23), 8614.
12. Estiller, J.; Bawag, D; and Martirez, B. (2018). Filipino-badwords-list. Retrieved May 15, 2021, from https://github.com/jromest/filipino-badwords-list/blob/master/src/filipino-badwords-list.js
13. Gholamy, A; Kreinovich, V.; Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. *Departmental Technical Reports* (*CS*), 1209.
14. Diaz, G. (2016). stopwords-tl. Retrieved May 15, 2021, from https://github.com/jromest/filipino-badwords-list/blob/master/src/filipino-badwords-list.js
15. Flores, H. (2021). Phippines records 51 degree celsius heat index in Pangasinan. Retrieved May 13, 2021, from https://www.philstar.com/nation/2021/05/10/2097066/philippines-records-51-degrees-celsius-heat-index-pangasinan