Nigerian Journal of Technology (NIJOTECH)
Vol. 41, No. 2, March, 2022, pp.311-317
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
http://dx.doi.org/10.4314/njt.v41i2.13

Hostile Social Media Harassment: A Machine Learning Framework for Filtering Anti-Female Jokes

I. I. James[1, *], V. I. Osubor[2],

[1]*Department of Computer Science, Akwa Ibom State University, Akwa Ibom State, NIGERIA*
[2]*Department of Computer Science, University of Benin, Edo State, NIGERIA*

Abstract
Drawing motivation from the non-existing computational criteria that blocks abusive contents targeted at female folks on social media from weird contents capable of hurting, offending, or intimidating the feelings; we present a machine learning framework for an intelligent filtering of information regarding anti-female jokes on social media discourse using multi-class classification. Our framework trains, validates and test neural network algorithm with data set sourced from chat rooms on social media handles with the goal of detecting and filtering expression(s) available in online jokes defined in terms of harsh, mild and neutral anti-female jokes. During implementation, the prototyped model recorded a performance accuracy of 92.9 %. This performance though high with the available limited data set is yet to attain the desired level of accuracy due to lack of consistent rules that distinguish various classes of anti-female jokes. In future, this model will be deployed to achieve unhindered discourse on social media by implementing an intelligent filtering system that could identify, classify and block unwanted contents.

Keywords: *Social Media Harassment, Hostility, Text Filtering, Anti-Female Jokes, Machine Learning, objectifying comments.*

1.0 INTRODUCTION

Social media has provided global networking platforms to individuals of different backgrounds, ethnicities and culture to interact and share ideas. The influence of these platforms has pervaded nearly every aspect of human endeavours and changes the way people interact and share ideas [1]. However, in recent times, social media spaces have been inundated with various forms of crude or objectifying comments, pornographic contents and derogatory jokes, meant to harass and molest innocent users of such platforms. These practices have rendered social media environment hostile to some users due to increasing menace of cyber-bullying in the form of rampant gender-based violence (such as: rape, sexual molestation, harassment and defilement), especially among women and vulnerable individuals. One of the ways which hostile media space is created is through uncontrolled posting of contents of obscene jokes that provokes hostility and aggression. Mitchell [2] posited that jokes telling could portray hostility and aggression in both content and context, and tailored towards women to cause sexiest hostility. To curb the menace of rampant gender-based violence among women and marginalized individuals on social media platforms, anti-female jokes could be identified, classified and blocked through intelligent information filtering mechanism.

One of the ways of achieving an intelligent filtering mechanism is by using machine learning approach. Given the availability of large corpus of weird contents on social media, machine learning algorithms (such as: Artificial Neural Network (ANN), Support Vector Machine (SVM), k- Nearest Neighbour (kNN) and so on) have proven their effectiveness in solving classification problems involving textual, image, multimedia and audio data in different application areas [3, 4]. This high performance learning algorithms have been adopted by studies on information filtering in recent times [5, 6]. However, a few studies extended to filter gender-based harassment and intimidation on social media environment often adopts traditional approach of data analysis to determine the validity of contents on social media. With this approach, low performance accuracy due to language bias and failure to define the targets of the abusive words [7] is recorded. Secondly, limited or non-availability of data set needed for adequate machine learning processes arising from low participation of minority groups on the social platform, poses a challenge

***Corresponding author (Tel:** +234 (0) 7030985358**)**
Email addresses: idarajames@aksu.edu.ng (I. I. James) an vosubor@yahoo.com (V.I. Osubor)

to accurate classification results of information. Lastly, the overall lack of precise formal definition of features on anti-female joke capable of providing exhaustive model training and validation also lead to classification error in the learning approach.

Therefore, this paper specifies the rules for digitizing features of anti-female jokes described in terms of humour-specific feature characteristic of one-liner jokes grouped as human-centeredness and polarity orientation features [8]. A neural network model is developed to classify sexiest humour information in terms of harsh, mild, or neutral using available data from anti-female jokes. This is with a view to build filtering model to curb hostilities of social media environment in terms of sexual harassment among female users of the platform. The main contribution of this paper is the creation of digitized features of anti-female jokes obtained without any specific software from chat feeds of the social media. Again, the digitized data is used to design an intelligent framework of anti-female information filtering on social media using ANN.

The remaining part of this paper is further organized as follows: In section 2.0, literature review is presented. Section 3.0 addresses the methodology and anti-female joke framework and Section 4.0 implements the information filtering components of the framework while section 5.0 presents the discussion of result and conclusion.

2.0 LITERATURE REVIEW

A number of researches on sexiest humour expressed on social media platforms have been studied theoretically for different intents of users of social media platforms, without pre-establish criteria for implementation on machine learning algorithm.

Strain et al. [9] studied users' behavior on Facebook to examine the perceptions of men and women using sexiest humour based on statistical technique to discover the benefits and setbacks posed by both genders. The study provided useful information to allow expansion of research on humour studies within the social media platforms. Nonetheless, it was a quantitative study on gender-based perception of sexiest humour on social media platform. Zhong et al. [10] studied cyber-bullying based on information posted on Instagram by its users, sharing images with a view to developing an early warning mechanism to detect weird contents. A Convolutional Neural Network (CNN) was adapted to learn the features of images posted on the platform with higher level of accuracy in performance. Gitariet et al. [11] developed a lexicon-based model that used sentiment analysis to classify hate speech in web-based discussion. Although semantic, hate and theme-based features were used for building the classification model; the definition of hate speech is ambiguous when it is generalized for every group of individual. Ji et al. [12] investigated how suicidal tendencies can be detected from the online user-generated contents using supervised learning approach. More so, Sawhney et al. [13] used deep-learning to explore and learn the suicidal ideation for accurate detection on tweets. This study was limited to classification of suicidal tendencies generated by the users of online media. Novalita et al. [14] also studied the identification of tweets with cyber-bullying and non-cyber-bullying intents using random forest. Although machine learning approach was adopted, the cyber-bullying contents defined the perceived defamatory attacks experienced by the users of social media without considering the weird contents targeted at specific groups of individual.

Maghfiroh & Muqoddam [15] studied the dynamics of sexual harassments on social media by conducting quantitative survey on social media. It was aimed at analyzing forms of sexual harassment and identified the factors responsible for such abuse. The result of the study revealed that sexual harassment on social media is directly or indirectly related to the meaning of the sentences posted by the users of the platform. This was a survey without the design of computational framework to further the research to the level of providing an automated system to detect sexual harassment on social media.

Park & Fung [16] studied both one and two-step classification techniques to detect abusive languages on Twitter and also classify same to specific types of abuses using Convolutional Neural Networks (CNN). The classification approach used character and word-level inputs obtained from data on sexiest and racist languages found on Twitter platform. The study though relevant to the study at hand failed to address the specific challenges of sexual harassment of women on social media which is portrayed in diverse texts, images, audio and videos. Lee et al. [17] performed comparative studies on different learning models for detection of hate and abusive speeches on Twitter. The study segmented the textual data with word-level features using n-gram ranging from 1 to 3, and character-level features from 3 to 8-grams. Each classifier was implemented using five machine learning classifiers.

Chandra et al [18] captured the structure of online and linguistic behaviours of online communities to detect abusive language using Graph Convolutional Networks (GCN). However, the study failed to address cyber-bullying of women in terms of sexual harassment over social media platforms.

Basak et al. [19] designed application for automating detection of public shaming on Twitter platform using Support Vector Machine (SVM). The study was limited to categorization and classification of shaming tweets but failed to address issues related to anti-female jokes.

Chen & Soo [20] proposed CNN with augmentation of filter sizes and numbers to develop humour recognition system using datasets with distinct joke types in both English and Chinese.

Weller & Seppi [21] proposed a model that learns to identify humorous jokes based on ratings gleaned from Reddit pages. However, ratings information is often unavailable as most users failed to respond to rating request.

Ibrohim & Budi [22] performed comparative analysis of machine learning models on abusive language and hate speech, including: detecting the target, category, and level of hate speech in Indonesian Twitter. The result presented revealed that Random Forest Decision Tree (RFDT) classifier using Label Power-set (LP) as the transformation method gives the best accuracy with fast computational time.

Based on the reviewed literature, it is evident that the trend of filtering humour-related features from social media is shifting towards machine learning approaches. However, there is dearth of literature on current machine learning techniques/algorithms to address the undaunted challenge of hostile social media space. Hence, this study proposes an intelligent approach targeted at female folks to curb the menace of cyber bullying and harassment.

2.1 Social Media

Social media is a tool that boosts communication, interaction and connection among different racial groups, gender, and families across the world [23]. Some examples of social media environment include: Facebook, Twitter, Instagram, and more. It is a tool that can open several opportunities for e-commerce, e-learning, e-banking, and e-government to thrive. This is achieved through advertisement, governance, provision for learning and research tools, sharing of ideas with broad audience, professional skills acquisition, among its users. These platforms are sometimes subjected to abuse by individuals with the intention of bullying, privacy invasion, harassments, shaming and others [23, 13, 19]. Therefore, there is the need to develop filtering mechanism to remove the constraints to performance efficiency and ensure support for detection of abusive contents on social media [24].

2.2 Anti-female jokes on Social Media

Sexual harassment over the Internet can hinder free, legitimate, functional and joyful use of online platforms by social media users. This practice often necessitates hostility of social media thus leading to emotional harm of the legitimate users of the online platform [19]. In social media platforms where meaningful discourse is initiated, sexual harassment prevails in the form of gender harassment and unwanted sexual attention expressed as humour of different types, namely: sexiest, sexual, racist, profanity, homophobic and shaming and so on [19, 15, 9, 25].

Billig [26] stated that jokes are 'just' jokes, or 'just' a clever play of form, and not the expression of problematic motives'. A study conducted by [8] described anti-female jokes in terms of humour-specific feature characteristic of one-liner jokes. These features are grouped as human-centeredness and polarity orientation [8]. Examples of human-centeredness include: human-centric vocabulary, professional communities and vulgar language. Similarly, examples of polarity orientation include: negation, negative orientation and human weakness. These features are represented in Figure 1.

Typical examples of anti-female jokes excerpts from social media are given in Joke I and Joke II as follows:

Joke I*: If women aren't supposed to be in the kitchen, then why do they have milk and eggs inside them?*

Joke II*: How do you get a dishwasher to dig a hole? Give the woman a shovel!*

From the aforementioned jokes, characteristic features in terms of human centeredness include: human-centric vocabulary (e. g. *women/ woman*), vulgar language is exemplified in *dishwater to dig hole and shovel, milk and eggs inside them*. Examples of polarity orientation include: *aren't, how, why and kitchen.*

3.0 THE PROPOSED FRAMEWORK

The proposed intelligent framework of anti-female jokes filtering is developed with ANN algorithm with a view to providing intelligent information filtering solution for achieving non-hostile social media space for female folks. In this framework, ANN is used to classify different types of anti-female jokes into any of its type (e.g. terms of harsh, mild and neutral), which is needed to filter weird contents capable of causing hostility among the social media users. The proposed framework of anti-female jokes filtering is shown in Figure 2. This framework is implemented with algorithm described in Algorithm 1.

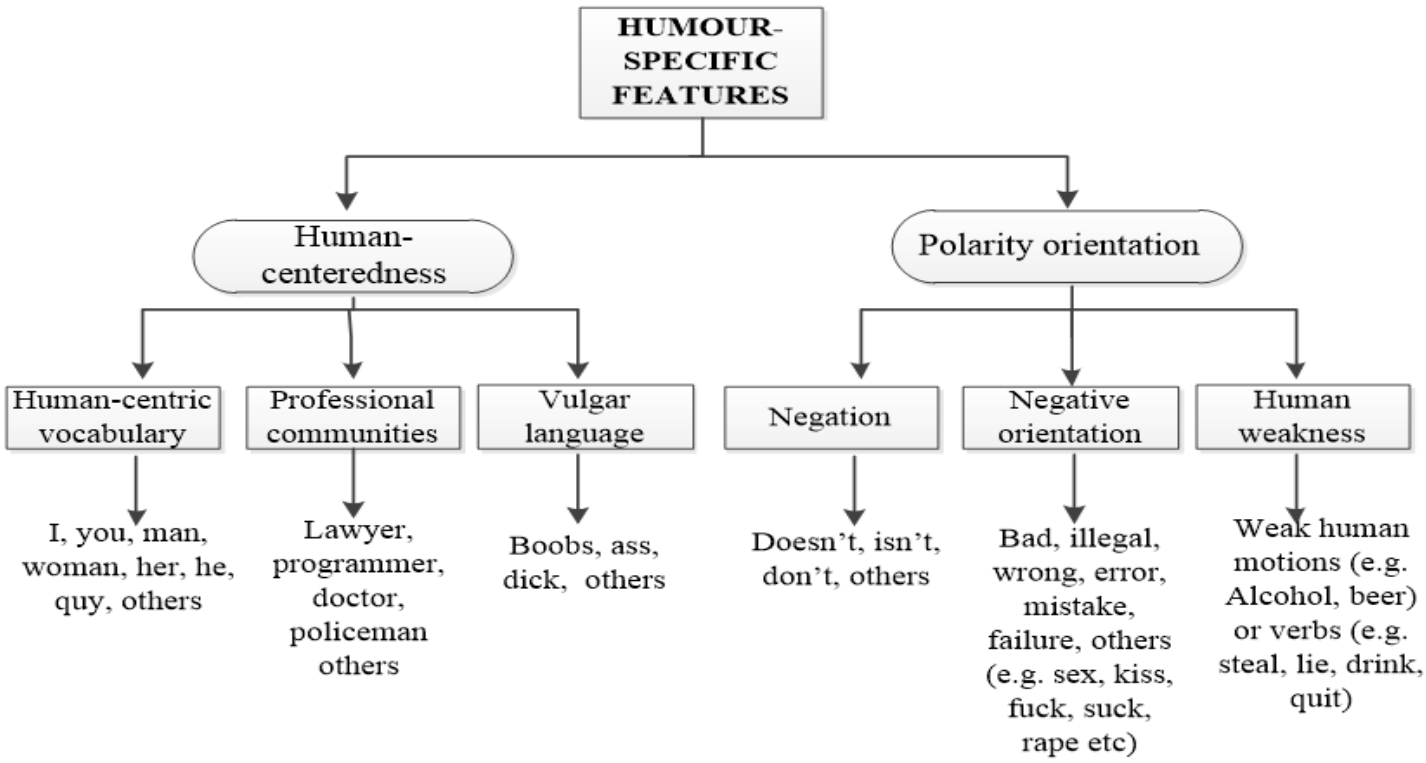

Figure 1: A Modified Representation of Humorous Features of Mihalcea and Pulman [8]

3.1 Description of the proposed framework

In the proposed framework, social media users' can generate text-based contents in the form of anti-female jokes. The generated contents are subjected to the machine learning module where the media contents are classified in three categories namely: harsh, mild and neutral. In order to achieve intelligent filtering of the media content, relevant features are extracted, trained, validated and tested to model the intelligent filtering module.

During the model implementation, outputs are compared with the predefined media contents that users should be allowed access to view. The approved contents are sent back to the users to view while unapproved contents are blocked. This framework is presented in Figure 2.

3.1.1 Algorithm 1: Algorithm of the Filtering Model

i. Extract frequency from anti-female jokes in terms of human centeredness and polarity orientation.
ii. Normalize the dataset between the ranges of (0-1) to avoid over fitting.
iii. Partition the data into k-fold for cross validation
iv. Generate the training and test dataset
v. Feed generated data into an ANN model for classification
vi. Evaluate the system with Matlab and test the performance with accuracy, precision and recall.

4.0 EXPERIMENT

The experiment was performed using data sets of sexist humour expressed in the form of one-liner anti-female jokes with characteristics attributes of human-centeredness and polarity orientation [8]. The binning of these features resulted in six joke features, namely: human-centric vocabulary, negation, negative orientation, sexiest terms, professional communities and private parts. Three hundred and twelve (312) one-liner humorous jokes comprising of 6 attributes each were sourced manually, since there was no automated tool available for extraction of one-liner humorous jokes from chat rooms of various social media handles, such as: Twitter, Instagram and Facebook.. It was also based on the definition of anti-female joke as reported in [8].

Relevant features from each one-liner humorous joke collected were extracted manually by experienced and trained annotators based on humor specific features to generate the frequency and likelihood measurements of

humorous features in anti-female jokes. A total of six (6) features were extracted for each definition of anti-female jokes, resulting in 1,872 features for digitization and pre-processing. Selected features of the dataset were transformed into suitable digitized format, to serve as input into machine learning module

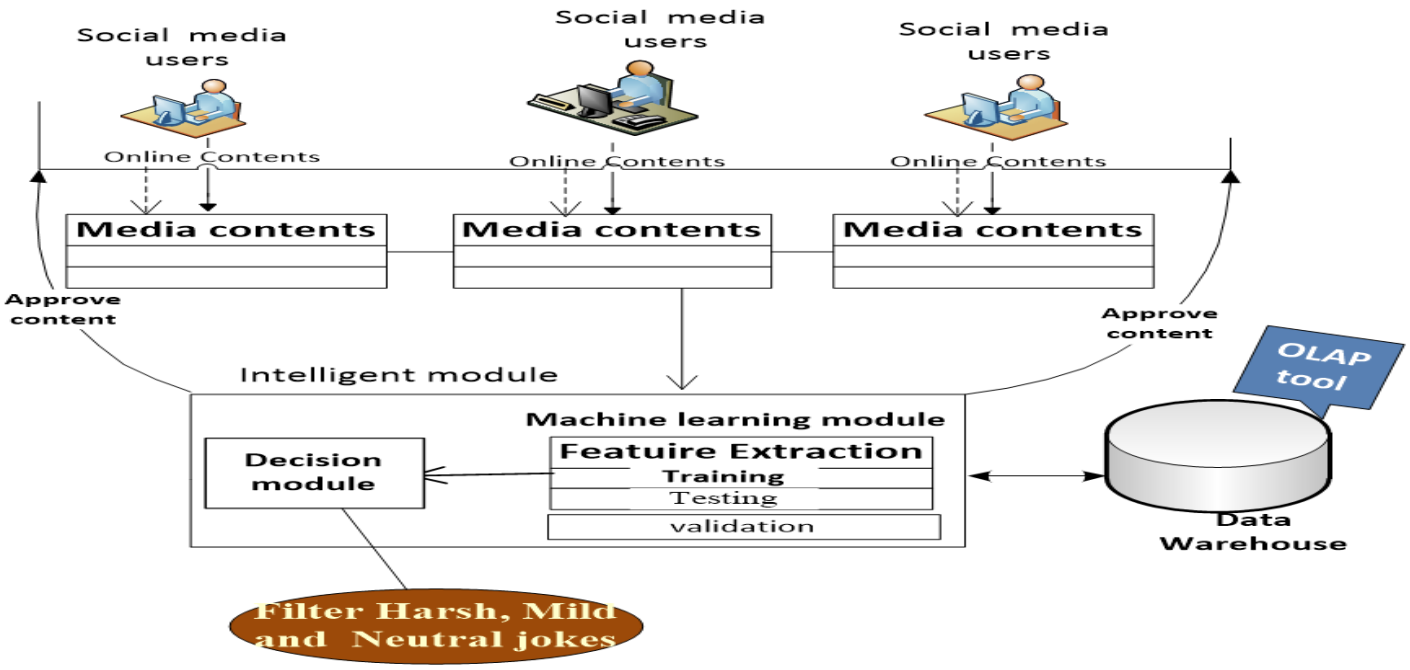

Figure 2: Framework of Intelligent Filtering of Anti-female Jokes

Distinctive pre-processing steps were employed to process the data before subjecting it to artificial neural network. It involves cleaning the input text to improve the accuracy of the proposed methodology, reducing the number of features and removing irrelevant, redundant or noisy data. Selected dataset was normalized for the learning process. The learning process adopts the basic steps of intelligent approach of model development used typical in Artificial Neural Network, which consists of data preprocessing, design and implementation using Matlab. The implementation adopted *k*-fold cross validation technique where data is split into *k*-equal parts (where *k*=10). In the first iteration, one part of data was selected as the test set and the model was trained on the remaining *k*-1 sets. Test error rate was then calculated after fitting the model to the test data. In the second iteration, the second set of data was selected as a test and the remaining (k-1) sets are used to train the data and determine its error. The above was repeated *k* times while changing the test part one-by-one until testing was completed on all the k-parts. Owing to small size of dataset, the ideal choice is *k*-fold cross validation with large value of k. 10-folds was selected because the higher the value of k, the higher the accuracy in cross-validation. To achieve quality classification for accurate filtering of anti-female jokes it is imperative to strive to achieve the highest accuracy, provided there is no trade-off in executable time.

5.0 RESULT AND DISCUSSION

A confusion matrix is a summary of prediction results on a classification problem. Based on the confusion matrix obtained from training, validation and testing of the developed model as shown in figure 3a and 3b, harsh, mild and neutral were represented with values of 1, 2, and 3, respectively for both outputs and target classes. Misclassified samples are presented in the red potions in the confusion matrix while green portions depict correctly classified samples.

During the training phase, the analysis of the result revealed that only 2 (1.0 %), 9 (4.7 %) and 150 (77.7 %) representing harsh, mild and neutral anti-female joke, respectively were correctly classified out of 193 available jokes. However, out of 150 (77.7 %) classified samples, 3 (1.6 %) and 25 (15.0 %) were misclassified as harsh and mild anti-female jokes respectively.

Furthermore, during validation, the analysis of the result revealed that only 1 (2.4 %) and 35 (83.3 %) representing mild and neutral anti-female joke were correctly classified out of 42 total available jokes. However, 1 (2.4 %) and 5 (11.9 %) out of 35 (83.3 %)

were misclassified as harsh and mild anti-female jokes respectively.

During the testing phase, performance accuracy of 92.9 % at 14 epochs was recorded as precision and 7.1 % as recall respectively. The analysis of the result revealed that only one harsh (2.4 %) anti-female joke was correctly classified out of 42 total available jokes. There was neither mild nor neutral anti-female joke on the test dataset. However, 38 (90.5%) and 3 (7.1%) was correctly and incorrectly classified as neutral anti-female jokes, respectively. From the results recorded in the confusion matrix of Figure 3a and 3b, it is obvious that there was insufficient test dataset to evaluate the accuracy of the intelligent classification model for filtering anti-female jokes. Nevertheless in future, this study would seek to improve on parsing of the accurate meaning of anti-female jokes as pointed out by [7] and source for large corpus of anti-female jokes. In addition, apart from the humorous features studied by [8], other meaningful features depicting weird and derogatory contents on online discourse targeted at other gender apart from female folks should be extracted to further enhance the filtering model.

6.0 CONCLUSION

Social media has provided global networking platforms to individuals but also serve as a tool for promoting sexual harassment of women in the form of crude or objectifying comments, pornographic contents and derogatory jokes. Therefore, this study adopted ANN to develop a framework for moderating the contents of social media to prevent hostile environment among its active users. The precision (92.9%) and recall (7.1%) obtained from the developed model showed that its weird contents targeted at female folks can be classified and filtered from the social media to promote healthy discourse among the users on the platform.

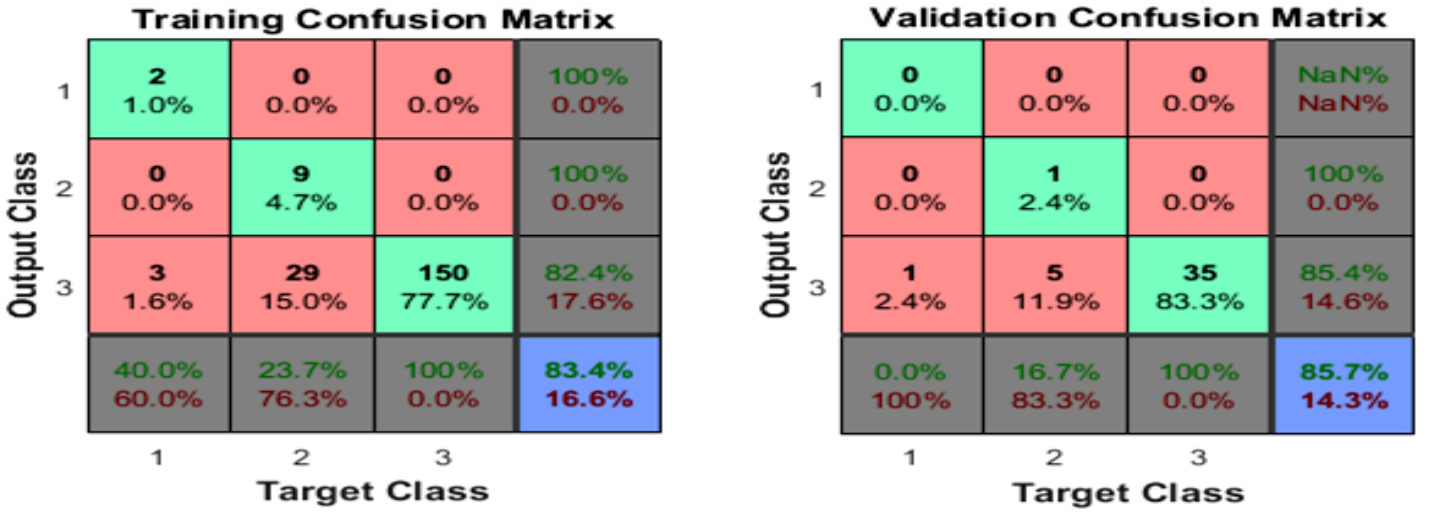

Figure 3a: Confusion matrix of model training and validation

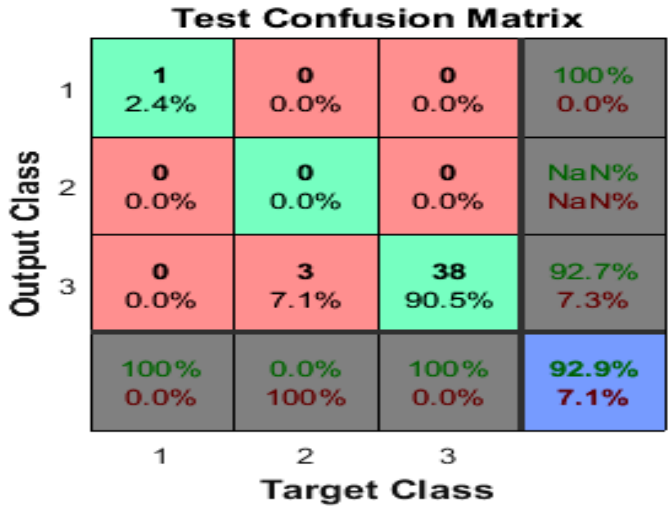

Figure 3b: Confusion matrix of model testing

REFERENCES

[1] Fang, B., Jia, Y., Han, Y., Li, S., and Zhou, B. "A survey of social network and information dissemination analysis", *Chinese science bulletin*, 59(32), 2014, pp. 4163-4172,

[2] Mitchell, C. A. "The differences between male and female joke telling as exemplified in a college community", Indiana University, 1976

[3] López-Úbeda, P., Díaz-Galiano, M. C., Martín-Noguerol, T., Luna, A., Ureña-López, L. A., and Martín-Valdivia, M. T. "Automatic medical protocol classification using machine learning approaches", *Computer Methods and Programs in Biomedicine, March* 2021.

[4] Ikonomakis, M., Kotsiantis, S., and Tampakas, V. "Text classification using machine learning

techniques", *World Scientific and Engineering Academy and Society Transactions on Computers*, 4(8), 2005, pp. 966-974.

[5] Heggo, I. A., and Abdelbaki, N. "Data-Driven Information Filtering Framework for Dynamically Hybrid Job Recommendation", *In Intelligent Systems in Big Data, Semantic Web and Machine Learning*, Springer, Cham. pp. 2021, pp. 23-49,

[6] Mouzannar, H., Rizk, Y., and Awad, M. "Damage Identification in Social Media Posts using Multimodal Deep Learning.", *Proceedings of the 15th Information Systems for Crisis Response and Management Conference – Rochester, NY, USA*, May 2018.

[7] Duarte, N., Llanso, E., and Loup, A. C. "Mixed Messages? The Limits of Automated Social Media Content Analysis". *Infant Journal*, January 2018, pp. 106.

[8] Mihalcea, R. and Pulman, S. "Characterizing Humour: An Exploration of Features in Humorous Texts", http://www.eecs.umich.edu/cse/awards/pdfs/mihalcea.cicling07.pdf. Accessed on October 4, 2021.

[9] Strain, M., Saucier, D., and Martens, A. "Sexist humor in Facebook profiles: Perceptions of humor targeting women & men. Humor", 28(1), . 2015, pp. 119-141. DOI 10.1515/humor-2014-0137

[10] Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., and Caragea, C. "Content-Driven Detection of Cyberbullying on the Instagram Social Network" *International Joint Conference on Artificial Intelligence*, Vol. 16, 2016, pp. 3952-3958.

[11] Gitari, N. D., Zuping, Z., Damien, H., and Long, J. "A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10, Number 4, 2015, pp. 215-230.

[12] Ji, S., Yu, C. P., Fung, S. F., Pan, S., and Long, G. "Supervised learning for suicidal ideation detection in online user content". Complexity, September 2018.

[13] Sawhney, R., Manchanda, P., Mathur, P., Shah, R., and Singh, R. "Exploring and learning suicidal ideation connotations on social media with deep learning", *In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, October 2019, pp. 167-175.

[14] Novalita, N., Herdiani, A., Lukmana, I., and Puspandari, D. "Cyberbullying identification on twitter using random forest classifier", *In Journal of Physics: Conference Series*, Vol. 1192, No. 1, 2019, pp. 012029.

[15] Maghfiroh, V. S., amd Muqoddam, F. "Dynamics of sexual harassment on social media", *In International Conference of Mental Health, Neuroscience, and Cyber-psychology* Fakultas Ilmu Pendidikan. 2019, pp. 154-162.

[16] Park, J. H., and Fung, P. "One-step and two-step classification for abusive language detection on twitter", *arXiv preprint arXiv:1706.01206.*, June 2017

[17] Lee, Y., Yoon, S., and Jung, K. Comparative studies of detecting abusive language on twitter, 2018, *arXiv preprint arXiv:1808.10245*

[18] Chandra, S., Mishra, P., Yannakoudakis, H., Nimishakavi, M., Saeidi, M., and Shutova, E. "Graph-based modeling of online communities for fake news detection", *arXiv preprint arXiv:2008.06274*, August 14, 2020.

[19] Basak, R., Sural, S., Ganguly, N., and Ghosh, S. K. "Online public shaming on Twitter: Detection, analysis, and mitigation" *Institute of Electrical and Electronics Engineers Transactions on Computational Social Systems"*, 6(2), 2019, pp. 208-220.

[20] Chen, P. Y., and Soo, V. W. "Humor recognition using deep learning", *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies"* 2018, 2, pp.113-117.

[21] Weller, O., and Seppi, K. "Humor detection: A transformer gets the last laugh", *arXiv preprint arXiv:1909.00252*, August 31, 2019

[22] Ibrohim, M. O., and Budi, I. "Multi-label hate speech and abusive language detection in Indonesian twitter", *In Proceedings of the Third Workshop on Abusive Language Online* , August 2019, pp. 46-57.

[23] Brewer, B., Cave, A., Massey, A., Vurdelija, A. and Freeman, J. "Cyber Bullying among Female college students: an exploratory study". *Califonian Journal of Health Promotion*, 12(1), 2012, pp. 40-51.

[24] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. "Challenges and frontiers in abusive content detection", *In Proceedings of the third workshop on abusive language online*, August 2019, pp. 80-93.

[25] Bennett-Alexander, D. D. "Hostile environment sexual harassment: A clearer view", *Labor Law Journal*, 42(3), 1991, 131.

[26] Billig M. "Laughter and ridicule: Towards a social critique of humour", *Sage*; 2005.