EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 4 | Number 5 | 2024

Filtering Anti-Female Joke on Social Media Space: Natural Language Processing Approach

James Idara, Ekong Anietie, Udoh Abigail and Udoeka Ifreke
Computer Science Department,
Akwa Ibom State University, Ikot Akpaden, Nigeria

ABSTRACT

The growing threat of abuse from obscene jokes and other types of objectifying content especially among women has caused harassment and created a hostile environment for some users of social media space. To reduce the rate of hostility, filtering, therefore, becomes necessary for checking uncontrolled posting of contents of obscene jokes. The primary objective of this paper is to develop an intelligent filtering system of anti-female jokes on social media space using Natural Language Processing. 1500 one-liner anti-female jokes were sourced from social media sites, and expressed with characteristics attributes of human-centeredness and polarity orientation. The binning of these attributes was centered on: human-centric vocabulary, negation, negative orientation, sexiest terms, professional communities and private parts. The applicable dataset was divided utilizing k-fold cross-validation for the training process. A filtering system was developed utilizing the algorithm that exhibited the highest level of accuracy. The model was developed in Python, employing various Natural Language Processing techniques. Its performance was assessed using metrics such as precision, recall, and F1-score to ensure evaluation of its effectiveness. Results of the experiments showed that Random Forest algorithm produced the best accuracy with 95.3%. Therefore, the model could be adopted for intelligent filtering of anti-female jokes on social media.

Keywords: Anti-female, Filtering, Jokes, Natural Language Processing, Machine Learning Algorithms

INTRODUCTION

Filtering is a procedure used in extracting or choosing relevant items from a large repository based on certain principles (Belkin & Croft, 1992). It plays an important role in a decision-making process. This procedure can be used in various fields of study such as data analysis, image and audio processing, medical imaging, image enhancement, image restoration, image compression, image segmentation, and so on. For example, in the fields of finance or economics, where large datasets can be challenging to analyze, filtering can help simplify complex data or information by breaking it down into smaller and more manageable sizes (Patel *et al.*, 2022). In Information retrieval, filtering can be used to prevent unauthorized access or restrict access to specific content. The study conducted by Yang *et al.* (2022) posited that filtering content can help prevent online harassment and abuse by blocking harmful content and messages; this can promote a more civil and respectful online discourse. Filtering is categorized into content-based (Joy & Pillai, 2022), collaborative, hybrid (Papadakis *et al.*, 2022), and rule-based collaborative (Bahri *et al.*, 2021). In the field of psychology, filtering jokes can ensure that the humor is appropriate and does not offend or harm individuals or groups. However, content filtering (Tambe *et al.*, 2021) can be used to block websites with inappropriate or malicious content, thus enhancing security for users' information. Billig (2005) in his study defined a joke as a clever play or form and not an expression of problematic motives. It is a third-person narrative, usually told in the present

tense, and consisting of text by the anonymous author and dialogue by the different characters presented by and acting in the joke (Attardo & Chabanne, 1992). Davies (2010) proposed that jokes with similar themes have various conceptions in different countries since the amusement of a joke is determined by the context in which it is delivered. However, the main reason a joke is told is to amuse the audience, sometimes harass the listeners, or ridiculed the narrator of the joke. Generally, the telling of jokes is often done by male folks but female jokes at times result in humor (Martin & Ford, 2018) thus leading to hostility. A joke is said to be hostile if it is used as a means to put down a person or a group of people, or if the joke contains derogatory references to people, ideologies, or cultures.

Although there are several types of jokes such as Observational, Anecdotal, Situational, Character, Wit and sarcasm, One-liner obscene jokes, Ironic, Deadpan, Farcical, Self-deprecating, Slapstick, Self-deprecating humor, Absurd or surreal, and Satire, (Brunvand, 1985) assume that obscene (offensive) jokes are the most popular jokes and make up the largest number of jokes told among adults. One-liner jokes are jokes that are short and hilarious and are usually made using one word or one sentence. They are commonly used jokes, especially by comedians and actors due to their comedic method which promotes part of their act. Offensive anti-female jokes are jokes that target or reflect hate or discrimination against women.

Nevertheless, uncontrolled publishing of offensive jokes against women incites aggression and digital violence thus resulting in the generation of hostile media spaces. To reduce the magnitude of digital violence and sexual harassment against women and marginalized individuals on social media spaces, the need for filtering and blocking derogatory (offensive) comments is pertinent. To achieve effective filtering of offensive comments on social media space, there is a great need for identification and classification of the jokes based on quantifying different classes of its data. This is to support planning for preventive action in terms of strategic measures in managing derogatory media content.

Classification of anti-female jokes can either be done using a parametric or non-parametric model. A parametric model is a traditional statistical approach while a non-parametric model is an intelligent-based approach (Duan *et al.,* 2019; Ekong *et al.,* 2022). Parametric provide a formal framework for estimating the parameters of the model and testing the goodness-of-fit of the model to the data. However, one limitation of parametric models is that they may not be flexible enough to capture complex relationships between variables and may be prone to model misspecification if the assumed functional form does not fit the data well. The non-parametric approach can learn from a large data set and make predictions or decisions based on the learning. These models are often built using complex algorithms that allow them to analyze and interpret vast amounts of data, recognize patterns, and make predictions with a high degree of accuracy. Non-parametric can generalize solutions using any of the machine learning techniques, such as; supervised, unsupervised, semi-supervised, and re-enforcement learning (Dangeti, 2017; Ang *et al.,* 2015).

However, machine learning techniques are the strength of Natural Language Processing (NLP) which trains models to understand and generate human language. NLP is a field of artificial intelligence that focuses on the interaction between computers and human language. It involves the study and development of algorithms and models that enable computers to understand, interpret, and generate human language in a meaningful way. Additionally, it supports classifiers to handle unstructured textual data and extract meaningful information, leading to improved accuracy, efficiency, and automation in data classification tasks. In this study, supervised machine learning algorithms will be adopted to filter weird content capable of causing hostility among social media users. This algorithm is suitable with labeled data before subjecting it to machining. Although several studies on anti-female jokes have been conducted (Weller & Seppi, 2019; Maghfiroh & Muqoddam, 2019; Novalita *et al.,* 2019;

James & Osubor, 2022; Mihalcea & Pulman, 2007; Chen & Soo, 2018), but these studies failed to address the problem associated with harassment of women on social media. Following the reviewed literature, it is evident that existing studies adopted a conventional supervised machine learning approach to filter anti-female jokes but failed to implement a filtering system that could identify, classify, filter, and block unwanted contents. Therefore, to fill the knowledge gap, a multiclass-based model is proposed to support the filtering of anti-female jokes using NLP technology.

METHODS

In this study, text-based content in the form of anti-female jokes generated by social media users was subjected to pre-processing using NLP technique. The media contents were classified into three categories namely: harsh, mild, and neutral. However, to realize precise filtering of the media contents, relevant features from one-liner antifemale jokes were extracted from several social media handles, preprocessed, trained, validated, and tested using relevant machine learning algorithms. Achieving accurate filtering requires a formidable architecture; this is presented in Figure 1.

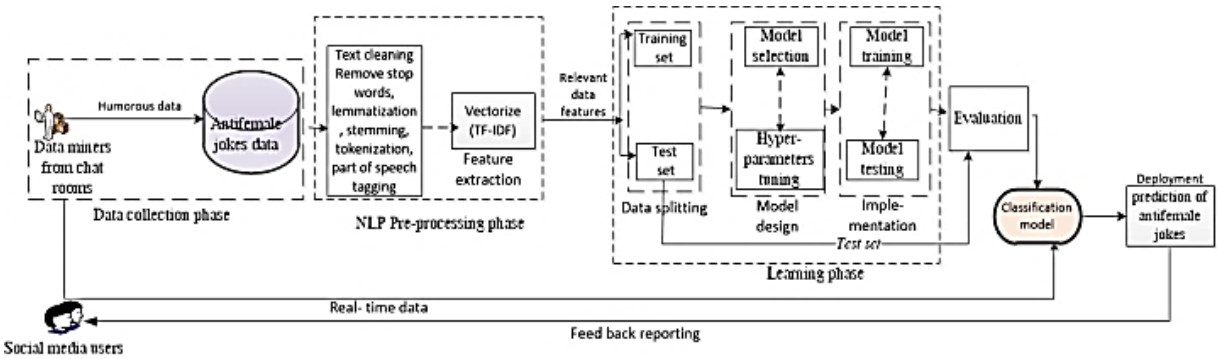

Figure 1: Proposed system architecture of precise filtering of anti-female jokes on social media space

Research Methods

The study employed the theory of predictive analytics (Kumar & Garg, 2018), which harnesses historical data to make informed predictions about future events. This methodology integrates machine learning, data mining, and statistical modeling to enhance its predictive capabilities. However, the proposed model was evaluated using datasets of sexist humor expressed in the form of one-liner anti-female jokes with characteristic attributes of human-centeredness and polarity orientation which was sourced from different social media sites such as Twitter, Instagram, Reddit, Snapchat, Facebook, Telegram, and WhatsApp.

In this work, 1,500 instances of data were sourced from different social media sites such as; Facebook, Instagram, Twitter, YouTube, Telegram, TikTok, etc. based on specific features of human-centeredness and polarity orientation. Relevant data features of human-centeredness and polarity orientation were extracted by trained annotators for pre-processing using NLP. The binning of these features resulted in six (6) features, such as human-centric vocabulary, professional communities, vulgar language, negation, negative orientation, and human weakness. For training, testing, and validation, the entire dataset was partitioned into k-fold and cross-validation procedures. The choice for k-fold cross-validation lies in the ability to split the dataset into equal parts for training and testing to achieve the highest accuracy (Yadav & Shukla, 2016). Suitable machine learning algorithms were used to learn the features obtained from the validation procedure. Confusion matrix was used to evaluate the performance of the predictive model. The model was evaluated with F1-score, Precision,

and Recall and implemented with the Python programming language. The raw data and modified representation of humorous features of anti-female jokes are as shown in Table 1 and Figure 2, respectively.

Table 1. Raw data of one-liner anti-female jokes

S/N	One-liner anti-female jokes
1.	how do you get a dish washer to dick a hole? give the woman a shovel
2.	if women aren't suppose to be in the kitchen, then why do they have milk and egg inside them?
3.	why did God give women yeast infections? So they would know what it's like to live with an irritating cunt for once
4.	What is the smartest thing that's ever come out of woman's mouth? Einstein's dick
5.	what do women and condoms have in common? They both spend more time in your wallet than on your dick
6.	you know if you really think about it, women are just like sandwiches, they taste alot better if no one else has fucked them
7.	what do women and parking spaces have in common? If all of the good ones are taken, stick it in the disabled one
8.	what is the useless skin around the vagina called? The woman
9.	what is the best thing to bring a first date with a woman ? Earplugs
10.	do you know how many genders there are? Just one because women are property
11.	One hundred women are not worth a single testicle
12.	what is the difference between women and terrorists? The terrorist can actually be negotiated with
13.	what is the difference between a woman and a baby? The baby will learn to grow up and stop complaining
14.	The idea that women belong in the kitchen is dead and offensive, i mean the rest of the house needs to be clean too
15.	Nature intended women to be our slaves. They are our property
16.	What do you call a woman with half a brain? Gifted

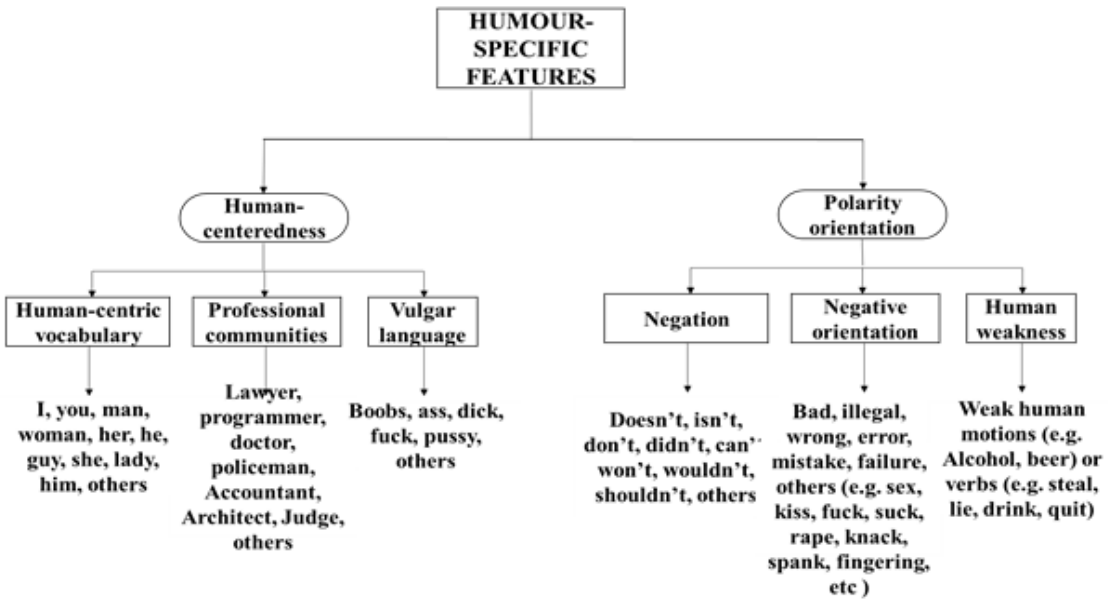

Figure 2: Modified representation of humorous features of anti-female jokes (Adapted from Mihalcea and Pulman, 2007)

Data Pre-processing Using NLP Techniques

The following steps were used for preprocessing of the data:

Text cleaning: texts were split into individual words or tokens and converted to lowercase to make the analysis case-sensitive, leading and trailing spaces were removed; regular expressions were used to remove them. HTML tags and different word forms were transformed into a common base, punctuation marks were replaced with spaces, and special

characters and digits were removed.

Text data discretization: raw text data was converted to a numeric vector; this serves as input for the machine learning algorithm. Some of the steps involved are:

Data transformation: Term Frequency-Inverse Document Frequency (TF-IDF) value for each term in the document was computed with the aim of converting the documents into TF-IDF feature vectors.

Data splitting: this was achieved using train_test_split; this function was used to split the dataset into training and testing subsets. The test size of 0.2 (20%) was used while 80% was used as the training set. The choice of the 80:20 ratio was to ensure the model generalized well on new data as well as improved accuracy and robustness. To obtain a more reliable estimate of model performance, utilize maximum data, model selection, and tuning, K-fold cross-validation was necessary where K=5. K-fold aid to mitigate the limitations of a single train-test split and provides a more comprehensive evaluation of the model's ability to generalize to any subsequent dataset. Also, the random state parameter was set to 42, which gave the best reproducibility of the splits.

Model design: suitable algorithms that suit the problem were selected; a loss function that measured the error between the predicted output and the actual output was selected. Furthermore, hyperparameters tuning such as the learning rate, batch size, and number of epochs were set during model training. The model was trained and tested using Random Forest, decision tree, artificial neural network, and k- nearest neighbor algorithms and evaluated using relevant validation parameters.

Model implementation: Miniconda toolkit in Python programming language was used for the implementation of the model. The best model accuracy was selected and used to develop the filtering system.

RESULTS

The result presented in Table 2 showed that the Decision tree (ID3) algorithm gave 94.29% accuracy, Artificial Neural Network (Multi-layer Perceptron Classifier) gave 89.26%, Random Forest (Random Forest Classifier) gave 95.30% and K-Nearest Neighbor (K-Neighbors Classifier) gave 74.16% of the model accuracy respectively. On comparing the results of the experiment, Random Forest produced the best result, with 95.30% accuracy, hence could be used in filtering anti-female jokes on social media.

Table 2: Discussion of result

Machine Learning Models	Mean cross-validation score	F1-score	Accuracy
Random Forest	0.937163776903167	0.9455047347614232	95.30%
Decision tree	0.9377583944970393	0.943077467476701	94.29%
ANN	0.8738502996136162	0.8858936117325379	89.26 %
K-nearest neighbor	0.7695493387228309	0.7164219273014751	74.16 %

DISCUSSION

The filtering system was developed in Python programming language and output is presented in Figures 3a, 3b, respectively. The snapshot results presented in Figures 3a, 3b showed the output of the filtering process based on multiclass classification of the 1500 one-liner jokes from different social media platforms. This proofs that our system could ultimately predict antifemale jokes based as neutral, mild or harsh jokes. Further description shows that if the joke is harsh, the filtering system automatically blocks the chat from the end user, if it is mild, the user gets a notification and if the joke is neutral, it means the joke is not harmful, hence may not result in harassment.

Figure 3a: Output of the filtering showing antifemale jokes "Neutral"

Figure 3b: Output of the filtering showing antifemale jokes "Harsh"

CONCLUSION AND RECOMMENDATIONS

The study aimed to reduce the rate of hostility among social media users by developing an anti-females joke filtering system. Humorous jokes data was collected and features were preprocessed using the NLP technique. The result of the preprocessing data was further subjected to machining. Relevant algorithms were selected for the filtering process leading to the implementation of a filtering system based on the performance of the most effective algorithm. However, through this initiative, the study seeks to contribute to the broader effort of curbing online aggression and promoting positive interactions in digital spaces. This initiative also has the potential to help mitigate hostilities directed towards the female gender within social media platforms.

Based on the aforementioned conclusion, the following recommendations are made:

- To ensure the model is inclusive and accurately identifies anti-female jokes, training data should be diverse and representative of different genders, cultures, and backgrounds.
- Social media platforms should collaborate with users and community groups to develop filtering strategies and incorporate feedback into the model's development.
- Additional dataset should be sourced for the purpose of further improving the model accuracy
- A combination of automated filtering systems, human moderation, user reporting mechanisms, and ongoing monitoring and updates is necessary.
- Continuous improvement, user feedback, and transparency in the filtering process can help address some of the challenges and create more effective and fair filtering systems.
- This research primarily relies on text analysis, classification of antifemale jokes can also be done using visual content, images, videos, and memes.

REFERENCES

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, *13*(5), 971-989. https://doi.org/10.1109/TCBB.2015.2478454

Attardo, S., & Chabanne, J. C. (1992). Jokes as a text type. *Humor*, *5*(1-2). https://doi.org/10.1515/humr.1992.5.1-2.165

Bahri, N., Bach Tobji, M. A., & Ben Yaghlane, B. (2021). ECFAR: A Rule-Based Collaborative Filtering System Dealing with Evidential Data. In *International Conference on Intelligent Systems Design and Applications* (pp. 944-955). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-96308-8_88

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin?. *Communications of the ACM*, *35*(12), 29-38. https://doi.org/10.1145/138859.138861

Billig, M. (2005). *Laughter and ridicule: Towards a social critique of humour*. Sage Publications.

Brunvand, J. H. (1985). Sex and Grammar Jokes. *New York Folklore*, *11*(1), 49.

Chen, P. Y., & Soo, V. W. (2018). 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* (Vol. 2, pp. 113-117). https://doi.org/10.18653/v1/N18-2018

Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing Ltd.

Davies, C. (2010). Jokes as the truth about Soviet socialism. *Folklore: Electronic Journal of Folklore*, *46*, 9-32.

Duan, S., Li, Y., Wan, Y., Wang, P., Wang, Z., & Li, N. (2019). Sensitivity analysis and classification algorithms comparison for underground target detection. *IEEE Access*, *7*, 116227-116246. https://doi.org/10.1109/ACCESS.2019.2936132

Ekong, A., Silas, A., & Inyang, S. (2022). A Machine Learning Approach for Prediction of Students' Admissibility for Post-Secondary Education using Artificial Neural Network. *International Journal of Computer Applications*, *184*, 44-49.

James, I. I., & Osubor, V. I. (2022). Hostile social media harassment: A machine learning framework for filtering anti-female jokes. *Nigerian Journal of Technology*, *41*(2), 311-317. https://doi.org/10.4314/njt.v41i2.13

Joy, J., & Pillai, R. V. G. (2022). Review and classification of content recommenders in E-learning environment. *Journal of King Saud University-Computer and Information Sciences*, *34*(9), 7670-7685. https://doi.org/10.1016/j.jksuci.2021.06.009

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. *International Journal of Computer Applications*, *182*(1), 31-37.

Maghfiroh, V. S., & Muqoddam, F. (2019, January). Dynamics of sexual harassment on social media. In *International Conference of Mental Health, Neuroscience, and Cyber-psychology* (pp. 154-162). Fakultas Ilmu Pendidikan. https://doi.org/10.32698/25272

Martin, R. A., & Ford, T. (2018). *The psychology of humor: An integrative approach*. Academic press.

Mihalcea, R., & Pulman, S. (2007). Characterizing humour: An exploration of features in humorous texts. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 337-347). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70939-8_30

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019, March). Cyberbullying identification on twitter using random forest classifier. In *Journal of physics:*

conference series (Vol. 1192, No. 1, p. 012029). IOP Publishing. https://doi.org/10.1088/1742-6596/1192/1/012029

Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., & Fragopoulou, P. (2022). Collaborative filtering recommender systems taxonomy. *Knowledge and Information Systems*, *64*(1), 35-74. https://doi.org/10.1007/s10115-021-01628-7

Patel, R., Goodell, J. W., Oriani, M. E., Paltrinieri, A., & Yarovaya, L. (2022). A bibliometric review of financial market integration literature. *International Review of Financial Analysis*, *80*, 102035. https://doi.org/10.1016/j.irfa.2022.102035

Tambe, U. S., Kakada, N. R., Suryawanshi, S. J., & Bhamre, S. S. (2021). *Content Filtering of Social Media Sites Using Machine Learning Techniques*. https://doi.org/10.3233/APC210226

Weller, O., & Seppi, K. (2019). *Humor detection: A transformer gets the last laugh*. arXiv preprint arXiv:1909.00252. https://doi.org/10.48550/arXiv.1909.00252

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International conference on advanced computing (IACC)* (pp. 78-83). IEEE. https://doi.org/10.1109/IACC.2016.25

Yang, J., Xiu, P., Sun, L., Ying, L., & Muthu, B. (2022). Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management*, *59*(1), 102751. https://doi.org/10.1016/j.ipm.2021.102751