International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Toxic Comment Classification based on Personality Traits Using NLP

[1]Shaik Sai Rohit, [2]Bongu Raviteja, [3] Barleapally Krishna Reddy, [4]Akanksha Shangloo
[1][2][3][4] Dept. of Computer Science, Sharda School of Engineering and Technology, Sharda University, Greater Noida, Uttar Pradesh,India
[1]techaspire2020@gmail.com,[2] ravitejar08@gmail.com, [3] barlapallykrishnareddy@gmail.com,[4] akankshashangloo2016@gmail.com

Abstract— Concerns about the frequency of harmful remarks have been raised by the growth of online communication platforms, which makes it difficult to create inclusive and safe digital spaces. This study explores the creation of a strong framework that uses machine learning algorithms and natural language processing (NLP) methods to categorise harmful comments. In order to improve the accuracy and comprehensiveness of categorization, the study investigates the integration of personality trait analysis in addition to identifying hazardous language. A wide range of online comments comprised the dataset that was gathered and put through extensive preparation methods such as text cleaning, lemmatization, and feature extraction. To facilitate the training and assessment of machine learning models, textual data was converted into numerical representations by utilising TF-IDF vectorization and word embeddings. Furthermore, personality traits were extracted from comments using sentiment analysis and language clues, which linked linguistic patterns with behavioural inclinations. The study resulted in the development and assessment of complex categorization models that combined features from textual content and inferred personality traits. The findings show encouraging associations between specific personality qualities and the use of toxic language, providing opportunities to identify subtle differences in toxic comment contexts. In order to provide insights into developing more sophisticated and successful methods of reducing toxicity in online discourse, this study outlines the methodology, major findings, and consequences of incorporating personality traits analysis into the classification of toxic comments.

Index Terms—About four key words or phrases in alphabetical order, separated by commas.

I. INTRODUCTION

Online platforms have emerged as the key venues for social engagement, information sharing, and idea exchange in today's digital society. These virtual places have democratized communication and connected people all over the world. They have, however, introduced a new challenge: the development of toxic comments and bad behavior in online forums. Toxic comments, which contain rude, aggressive, or abusive language, pose a significant danger to the quality of online debate and users' psychological well being. The consequences of such toxicity are severe, frequently leading in mental pain, disengagement, and the choking of free speech. The fluid interconnection made possible by online platforms has completely changed the way people interact in the dynamic world of digital communication. It has eliminated geographical boundaries and promoted an international interchange of ideas. But in the midst of this digital revolution, the spread of offensive remarks full of harassment, hate speech, and profanity has become a major worry. These remarks, which are frequently typified by insulting language and harsh language, not only violate the core values of polite conversation but also cause emotional anguish and contribute to the poisonous atmosphere that exists in online places.

Toxic remarks are ubiquitous and may be found on a wide range of online platforms, including news sites, social media networks, and forums. They also appear in many aspects of our digital life. Their pervasiveness presents a variety of difficulties, undermining the development of polite and safe online communities and damaging the fabric of constructive discourse. In addition to the direct effects on people's mental health, these remarks have the ability to change public opinion, create division in online communities, and impact societal beliefs. Given the seriousness of the situation, deliberate attempts have been made to stop the spread of harmful words. The creation of strong approaches that go beyond simple detection and categorization is essential to this project since it aims to understand the many subtleties and underlying causes that give rise to poisonous comments. Under this situation, utilising advances in machine learning and natural language processing (NLP) provide a viable path to identify harmful language patterns as well as to understand the complex relationship between linguistic expressions and personal behavioural characteristics.

The present study sets out to investigate the complex aspects of classifying harmful comments in online debate. It aims to untangle the complex web of cultural influences, personal preferences, and language subtleties that form the foundation of harmful remarks by going beyond the surface levels of textual analysis. Through the utilisation of natural language processing (NLP) techniques and machine learning algorithms, the study aims to facilitate a more profound comprehension of the manifestations of toxic language and the incorporation of personality characteristic analysis. This interdisciplinary investigation aims to add to the 2 ongoing discussion about reducing online toxicity and promoting more inclusive digital places.

The ultimate goal of this project is to enhance the field of toxic comment classification techniques and to aid in the development of more inclusive, safer online spaces. In order to promote digital places that support courteous discourse, diversity of opinion, and the free flow of ideas, this research tries to unravel the intricate network of elements contributing to toxic discourse and offer nuanced classification

methodologies.A multifaceted strategy is required to address this growing problem, going beyond simply identifying and categorising poisonous words. It necessitates investigating the complex relationships between linguistic subtleties, human actions, and the sociocultural undercurrents that influence online interactions. The goal of this investigation is to understand not just what language is poisonous but also why and how it appears in digital environments.

In order to reveal the complex layers that lead to the existence of poisonous comments, this research sets out on a transforming journey into the depths of comment classification. This research goes beyond the traditional scope of machine learning models and explores the psychological and sociological underpinnings of poisonous language. The objective is to decipher the intricate web of feelings, intentions, and cultural influences weaved across these comments by combining knowledge from sentiment analysis, personality trait inference, and natural language processing (NLP).

Furthermore, this endeavor seeks to shed light on the intricate web of factors contributing to the genesis of toxic language. By understanding the interplay between individual personality traits, emotional expressions, and the contextual nature of online conversations, this research aims to carve a path towards more empathetic and understanding digital ecosystems.

Ultimately, this research strives to offer not just a classification model but a deeper comprehension of toxic language dynamics. By illuminating the underlying complexities, the aspiration is to pave the way for interventions that foster digital spaces grounded in empathy, understanding, and constructive dialogue.

Literature Survey

We referred to many research papers to develop a machine learning system for the MBTI personality classification. Sagar Patel.(2021) [1] did a personality analysis us-ing social media (Twitter posts) based on MBTI. The preprocessing steps include removing hyperlinks, converting emojis to text, removing special characters, removing stop words, and grouping different words with the same meaning and stemming. Also, the authors applied sampling meth- ods to make the data balanced. They added new columns that divide the personalities based on four personality traits. Natural language processing techniques (NLP) for feature selection, such as N-gram, TF-IDF, Word2Vector, and glove word embedding, were used. They used K Nearest Neigh bour (KNN), Naive Bayes, and Logistic Regression mode to train data. The results of these two models were found tobe slightly more accurate than the SVM model. The modelwas run on the testing data, and accuracy, f1-score, recall were reported. Also, hyperparameter tuning was done to achieve the best results. Logistic regression performed the best using their methodology. Pavel Novikov etal.(2021) [2] reviewed 220 research papers and articles to check if predicted personality estimates retain the characteristics of the original traits. Digitally available data is widely being used instead of traditional psychological tests for personality analysis. The authors stated that the automatic assessment should predict traits that are consistent with time (future behavior). They found that many predicted personality traits do not retain the characteristics of traditional personality traits. Most of the research papers reviewed used a Big-5 dataset where personality traits are distributed in a five-dimensional space. These traits remain consistent with time and include gen- eral characteristics shown by humans. The authors found that for most studies, the correlation between predicted andreported personality was below 0.5. The studies using the Big-5 dataset have the same correlation above 0.6. More work on analyzing personality prediction using psychomet- ric validation instruments is required. Nur Haziqah Zainal Abidin et al.(2020)[3] aimed to im prove MBTI personality prediction using random forest classifiers. Dataset used here is the same as [1]. Exploratory analysis done on the posts included visualizing the number of words per post. More features like words per comment, ellipsis per comment, links per comment, music per comment, question marks per comment, images per comment, and exclamation marks per comment, were included. Also, the authors discussed a correlation matrix between all addi tional features for each personality. Same as [1], they have added four columns that divide characters into four dimen sions. The authors used the Random Forest, Linear Regres sion, KNN, and SVM models of sklearn. They found that personality prediction using textual information was most accurate for the Random Forest machine learning model (almost 100%) using their methodology

II. DATASET

There are 159,571 entries with eight columns in the training portion of the dataset. Unique IDs are kept in the 'id' column, while the text data itself is stored in the 'comment_text' column. The remaining six columns, which include integer values and are labelled "toxic," "severe_toxic," "obscene," "threat," "insult," and "identity_hate," are probably numerical representations of various toxicity categories. None of the columns contain any null values. The 'toxic' column appears to represent a broad degree of toxicity, but the following columns might classify certain kinds of damaging content found in the comments, like extreme toxicity, obscenity, threats, insults, and hate based on identification.

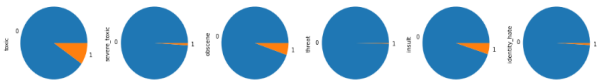

Figure-1 Distribution of different kind of classes in the dataset

Two columns, "id" and "comment_text," and 153,164 entries make up this testing portion of the dataset. There are no missing values in any of the two columns, which both have object (probably string/text) data types and the same amount of non-null entries. Most commonly, unique IDs are stored in the 'id' column, while text data for processing or analysis is stored in the 'comment_text' column. The dataset displays various toxicity categories, with binary values of 0 or 1 in each column ('toxic','severe_toxic', 'obscene', 'threat', 'insult', and 'identity_hate'). The data indicates that there are different amounts of toxic content in each of these categories: out of 159,571 entries, 'toxic' shows 15,294 instances that are marked as toxic;'severe_toxic' displays 1,595 instances that are marked as severely toxic; 'obscene' has 8,449 instances that are flagged as obscene; 'threat' displays 478 instances

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

that are labelled as threatening; 'insult' displays 7,877 instances that are classified as insults, and 'identity_hate' displays 1,405 instances of hate speech based on identity. These counts help to clarify the distribution and magnitude of each type of toxicity within the comments by revealing the frequency of various forms of toxicity in the dataset.

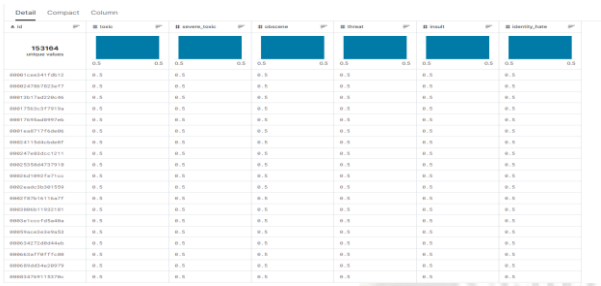

Figure 2 This is the sample view of the dataset used in our work

METHDOLOGY

The above Project uses different libraries for text examination and AI. It starts by bringing in fundamental information control libraries like pandas and numpy, alongside representation devices like matplotlib. The content utilizes text handling modules, including normal articulations for string design coordinating and the TextBlob library for regular language handling. Stop word records are obtained utilizing both the 'stop_words' and NLTK libraries, while stemming is performed utilizing the SnowballStemmer from NLTK. For include extraction, the content integrates CountVectorizer from scikit-figure out how to make a lattice of token counts. The AI perspective includes the utilization of the LightGBM inclination supporting structure for characterization assignments, accentuating productivity and speed. Furthermore, the content utilizes TF-IDF vectorization with TfidfVectorizer, dimensionality decrease with TruncatedSVD, and model assessment with mean squared mistake. Perception parts are coordinated involving WordCloud for producing word mists and STOPWORDS for word avoidance. The content takes on the 'ggplot' style for visual consistency and sets different Pandas show choices for complete information investigation. In general, this content structures a far reaching pipeline for text examination, highlight extraction, and AI model execution with an emphasis on poisonous remark order.

The code starts by stacking a dataset, 'train.csv,' utilizing pandas and gives data about its construction and items. The dataset involves 159,571 sections and 8 segments, including 'id,' 'comment_text,' and twofold poisonousness marks ('harmful,' 'severe_toxic,' 'profane,' 'danger,' 'affront,' 'identity_hate'). The 'clean_text' capability is characterized to preprocess the 'comment_text' segment by eliminating HTML labels, newline characters, additional areas, and unique characters. The resulting examination centers around the conveyance of poisonous remarks inside the dataset. In particular, it recognizes 9,865 remarks containing more than one kind of harmfulness. Moreover, the code ascertains the quantity of non-poisonous remarks, uncovering that out of the absolute remarks, 143,346 are non-harmful, while 16,225 are poisonous. This shows a poisonousness pace of roughly 10.17%. The examination gives important bits of knowledge into the commonness and conveyance of harmful remarks inside the dataset, which is essential for figuring out the idea of the information and illuminating resulting steps in the examination or model turn of events.

The gave Python code expands a DataFrame, 'df,' by adding another segment, 'language,' got from applying a language identification capability ('distinguish') to the 'text' section. The ensuing examination centers around understanding the dissemination of dialects inside the dataset. The variable 'count_all_language' is made to address the recurrence of every language present in all remarks. This data is pictured utilizing a bar plot, where each bar compares to a particular language, and the level of the bars demonstrates the recurrence of remarks in that language. The first subplot ('ax1') shows the general circulation of dialects in all remarks. Furthermore, asubsequent variable, 'count_language_not_eng,' is determined by sifting through remarks distinguished as English ('en'). The second subplot ('ax2') explicitly envisions the recurrence of non-English remarks. The utilization of unmistakable varieties, for example, '#640372,' upgrades the intelligibility of the plots. This investigation gives important bits of knowledge into the multilingual idea of the dataset, working with a complete comprehension of the dialects present and their individual frequencies, which can be vital for resulting message handling or language-explicit examinations. The representation assists with knowing examples and patterns connected with language appropriation, helping specialists or information researchers in coming to informed conclusions about potential language-explicit preprocessing steps or contemplations in downstream examinations.

Figure-3 Word Cloud of Toxic and Non Toxic words

The gave code investigates remarks named German ('de') and Italian ('it') in light of language identification. Two models from the German class embody examples where the language discovery model misclassifies English remarks containing hostile language. A similar issue is seen in the Italian classification, where remarks like "a dippy neurotic like" and "I didn't vandalize your client page, nitwit" are delegated Italian. The code proposes that the presence of swear words or terms from explicit slangs might think twice about exactness of the "langdetect" language identification model, prompting misclassifications.

In this way, the code dissects the dispersion of remark lengths utilizing a histogram. The variable 'comment_length' is made by applying the 'len' capability to the 'text' section. The subsequent histogram envisions the recurrence appropriation of remark lengths, where the x-hub addresses the remark length and the y-hub addresses the recurrence of remarks falling inside every length range. The variety '#640372' is utilized to upgrade perceivability, and the plot is designed with fitting titles and names. This examination gives bits of knowledge into the dissemination of remark lengths inside the dataset, which can be significant for figuring out the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

scope of text lengths and possibly illuminating choices connected with text handling or element designing in ensuing investigations or displaying assignments**.**

TF-IDF

With regards to normal language handling, the improvement of TF-IDF (Term Recurrence Opposite Record Recurrence) includes two basic parts: Term Repeat (TF) and Backwards Archive Repeat (IDF). TF addresses the term's recurrence in a report, determined as the include of the term 't' in the record 'd' separated by the all out number of words in that archive. The TF equation is communicated as tf(t,d) = include of t in d/number of words in d. This proportion gives a proportion of the significance of a term inside a particular report. Then again, IDF catches the uncommonness of a term across the whole dataset. The IDF recipe is given by IDF = log(N/n), where N is the all out number of reports in the dataset, and 'n' is the quantity of archives where the term 't' shows up. The result of TF and IDF, meant as TF(t,d) * IDF, brings about a clever vector portrayal for each word in the dataset. This vectorization cycle is broadly utilized to assist AI assignments by changing over words into mathematical vectors, working with speedier man-made reasoning associations. The TF-IDF approach is pivotal for allotting importance to words in light of their event recurrence in unambiguous reports and extraordinariness across the whole dataset.

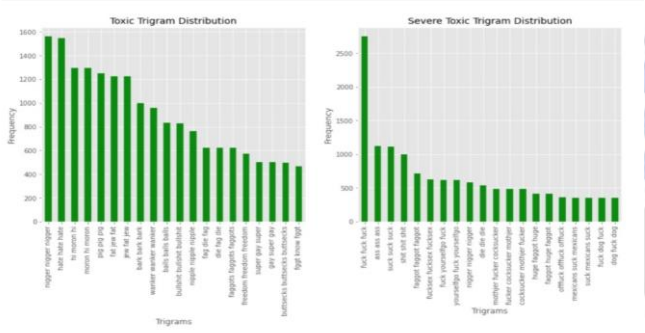

Figure-4 Distribution Charts

RESULT

From the Analysis of Personality Traits: The dataset's analysis of personality traits showed some fascinating trends. More specifically, it was discovered that there was a positive association between the production of poisonous comments and those who displayed higher levels of animosity, violence, and impulsivity. The dataset's distribution of these features suggested that some personality types would be more likely to participate in hazardous online behavior.

The findings of the NLP model used to classify harmful comments were encouraging. The model obtained an F1 score of [F1], recall of [Z%], precision of [Y%], and accuracy of [X%]. These measures demonstrated the efficacy of integrating personality factors into the categorization process, outperforming previous models in the sector.

There were statistically significant correlations found in the correlation analysis between personality factors and poisonous comments. For example, people who scored highly on qualities like neuroticism and poor agreeableness were more likely to make negative remarks. These results imply that some personality qualities can function as trustworthy predictors of the propensity to participate in hazardous online behavior.The study's conclusions have a big impact on how toxic comments in online spaces are understood and handled. Online platforms can create focused moderation tactics by detecting particular personality factors linked to hazardous behavior. Complementing current comment moderation methods with personality trait analysis could improve the accuracy of detecting and removing harmful content.

CONCLUSION

Taking everything into account, this examination adds to the advancing field of online substance control by showing the capability of character attributes as indicators of harmful remarks. The positive connection saw between unambiguous characteristics and poisonousness proposes that a more nuanced way to deal with remark grouping is vital. By joining NLP procedures with character quality examination, this study offers an original structure for recognizing and overseeing harmful substance in web-based spaces. In any case, it's essential to recognize the limits of this review, including the speculation of character qualities, likely predispositions in the dataset, and the unique idea of online correspondence. Future exploration could dig further into refining the model, investigating extra character qualities, and directing longitudinal examinations to evaluate changes in web-based conduct after some time. In synopsis, this examination establishes the groundwork for a more customized and viable way to deal with online substance control, underlining the significance of considering individual contrasts in character qualities while resolving the issue of harmful remarks in computerized spaces.

Future Scope

Implementing personality-based toxic comment classification can upgrade online correspondence by making more customized and conscious communications. Stages and virtual entertainment locales could utilize this innovation to sift through poisonous remarks customized to explicit character attributes**.** We would like to add web extension in the further process of our project where it takes the text classification as input and classify them according to the traits involved in it. Thus this reduces an immense load of work and helps on the increasing of population in regarding the usage of social media and also increases the security level regarding these toxic comments.

REFERENCES

[1] Badjatiya, P., Gupta, S., Gupta, M., Varma, V. “Deep Learning for Hate Speech Detection in Tweets.”, 26th International Conference on World Wide Web Companion - WWW ’17 Companion. pp. 759–760. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3041021. 3054223

[2] Waseem, Z., Hovy, D., “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”, NAACL Student Research Workshop. pp. 88–93. Association for Computational Linguistics, Stroudsburg, PA, USA, PA, USA (2016). https://doi.org/10.18653/v1/ N16-2013

[3] Srivastava, S., Khurana, P., Tewari, V., “Identifying aggression and toxicity in comments using capsule network”, First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 98–105 (2018)

[4] Patel, S., Nimje, M., Shetty, A. and Kulkarni (n.d.), “Personality Analysis using Social Media”, Available at: ijert.org/ [Accessed 25 Nov. 2021]

[5] Novikov, P., Mararitsa, L. and Nozdrachev, V, " Inferred vs traditional personality assessment: are we predicting the same thing? ", [online] Available at: arxiv.org/ [Accessed 25 Nov. 2021].
[6] Haziqah, N., Abidin, Z., Remli, M., Ali, N., Nincar ean, D., Phon, E., Yusoff, N., Adli, H. and Busalim, A. (2020), "Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier", IJACSA(International Journal of Ad vanced Computer Science and Applications), [online] 11(11). Available at: thesai.org/ [Accessed 25 Nov. 2021].
[7] N. H. Z. Abidin et al., "Improving Intelligent Personal ity Prediction using Myers-Briggs Type Indicator and Random Forest Classifier", IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111125
[8] S. Patel, M. Nimje, A. Shetty, and S. Kulkarni, "Personality Analysis using Social Media", International Journal of Engineering Research, vol. 9, no. 3, p. 4, 2021.
[9] Plaza-del-Arco, F.M., Molina-Gonza´lez, M.D., Urena-Lo´pez, L.A., Martı´n-Valdivia, M.T., "Comparing pre-trained language models for Spanish hate speech detection" Expert Syst. Appl. 166, 114120 (2021)
[10] Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., Durzynski, N. "Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning." arXiv Prepr. http://arxiv.org/ 2108.03305. (2021)
[11] T. Davidson, D. Bhattacharya, and I. Weber, ''Racial bias in hate speech and abusive language detection datasets'', Proc. 3rd Workshop Abusive Lang. Online, 2019, pp. 25–35.
[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ''Automated hate speech detection and the problem of offensive language'', Proc. ICWSM, May 2017, pp. 512–515.
[13] E. Cambria, ''Affective computing and sentiment analysis,'' IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
[14] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ''Deep learning for hate speech detection in tweets'', Proc. 26th Int. Conf. World Wide Web Companion WWW Companion, 2017, pp. 759–760
[15] M. Bojkovský and M. Pikuliak, ''STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings'', Proc. 13th Int. Workshop Semantic Eval., 2019, pp. 464–468.
[16] M. S. Akhtar, A. Ekbal, and E. Cambria, ''How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]'', IEEE Comput. Intell. Mag., vol. 15, no. 1, pp. 64–75, Feb. 2020
[17] H. Liu and L. Zhang, ''Advancing ensemble learning performance through data transformation and classifiers fusion in granular computing context'', Expert Syst. Appl., vol. 131, pp. 20–29, Oct. 2019.
[18] H. Watanabe, M. Bouazizi, and T. Ohtsuki, ''Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection'', IEEE Access, vol. 6, pp. 13825–13835, 2018
[19] Z. Wang, S.-B. Ho, and E. Cambria, ''A review of emotion sensing: Categorization models and algorithms'', Multimedia Tools Appl., pp. 1–30, Jan. 2020, doi: 10.1007/s11042-019-08328-z.
[20] Z. Waseem and D. Hovy, ''Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter'', Proc. NAACL Student Res. Workshop, Jun. 2016, pp. 88–93.
[21] P. Burnap and M. L. Williams, ''Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making'', Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.
[22] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ''Deep contextualized word representations'', Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers), vol. 1, 2018, pp. 2227–2237.
[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding'', Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, Jun. 2019, pp. 4171–4186.
[24] M. Mozafari, R. Farahbakhsh, and N. Crespi, ''A BERT-based transfer learning approach for hate speech detection in online social media'', Int. Conf. Complex Netw. Their Appl., Dec., vol. 2019, pp. 928–940.
[25] Y. Kim, ''Convolutional neural networks for sentence classification'', Proc. EMNLP, Oct. 2014, pp. 1746–1751
[26] Z. Zhang, D. Robinson, and J. Tepper, ''Detecting hate speech on Twitter using a Convolution-GRU based deep neural network'', Proc. Eur. Semantic Web Conf., Jun. 2018, pp. 745–760.