iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: VII Month of publication: July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36860

Toxic Comment Classification

Sonika Prakash[1], Poornima G B[2], Spoorthy R S[3], Anu Bai[4], Ramyashree G T[5], Nirmala C R[6]
[1, 2, 3, 4, 5, 6]*Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, VTU University*

Abstract: A large proportion of online comments available on public domains are usually constructive. However, a significant proportion is toxic and destructive. Several platforms and social media sites are finding it difficult to maintain fair conversation and are often forced to either limit the user comments or get dissolved by shutting down user comments completely. So, to prevent these types of identity hate through comments on social media, we come up with a solution to detect different types of toxicity in the comments using Deep Learning and Natural Language Processing. Dataset is obtained online which is processed to remove noise. Transformation of raw comments is done before feeding it to the classification model using Natural Language Processing. A Convolutional Neural Network model is used which will differentiate toxic comments from non-toxic comments.

Keywords: online comments; social media; toxic comments; deep learning; natural language processing; convolutional neural network

I. INTRODUCTION

Nowadays, billions of users actively using social networking platforms such as Reddit, YouTube, and Facebook. The popularity of Internet-based discussions has led to greater anonymity in these conversations. In most of these online discussions, toxic and hatred environments emerge. Exploring the interactions that cause these toxic environments to develop is important to eliminate this problem. In addition, understanding the toxic “composition” of an environment, such as the relative frequencies of obscenities, insults, threats, and comments propagating identity hate can also provide insight into how to fight these issues. Our system aims to understand the characteristics about these toxic conversations using deep neural networks. The continuous use of hateful or offensive language online has been growing rapidly in recent years, and the problem is now huge.

In some cases, toxic comments online have even resulted in real-life violence, from religious nationalism to neo-Nazi propaganda. Social media platforms which rely on the human reviewers, are struggling to reduce the ever-increasing volume of toxic content. It is known that Facebook is a place of repeated exposure to such toxic and distressing content. Bringing this work to machine learning can help moderate the rising volumes of toxic content while limiting user exposure to it. Detecting and controlling verbal abuse in an automated fashion is inherently a natural language processing task. It deals with the interaction between computers and humans using natural language.

The main objective of NLP is to read, comprehend, understand, and make sense of human languages. Most of these NLP techniques depend on machine learning to derive meaning from human languages [1]. Machine learning explores the development and study of algorithms that can learn and make predictions on data [2].

Such algorithms work by developing a model for test inputs to make data-driven predictions or decisions, rather than following strictly static program instructions. Toxic comment classification on online sites is conventionally carried out either by moderators or with the help of text classification tools [3].

With recent advances in Deep Learning (DL) techniques, researchers are exploring if DL can be used for the comment classification task. Text classification is a classic topic for natural language processing and an essential component in many applications, such as web searching, information filtering, topic categorization, and sentiment analysis [4]. Text transformation is the very first step in any form of text classification. Most of the online comments are usually in non-standard English and contain lots of spelling mistakes partly because of typos (resulting from small screens of the mobile devices) but more importantly because of the deliberate attempt to write the toxic and abusive comments in creative ways to dodge the filters.

II. LITERATURE REVIEW

Aggression by text is a complex phenomenon, and different knowledge fields try to study and tackle this problem. This analysis of related work focuses on a computer science perspective of aggression identification, a recently emerging area. Currently, the scientific study of automatic identification of aggressive text, using information technology techniques, is increasing. In this study, several related pieces of literature are used to express different types of aggression. Some of those are hate [8], cyberbullying [9], abusive language [3], toxicity [10], flaming [11], extremism [12], radicalization [4], and hate speech [13].

Despite the differences between those concepts, previous research can give us insight into methods to approach the problem of identifying aggressive interactions.

Focus is on the automatic detection of hate speech. For example, Georgakopoulos et al. [13] provide a short, comprehensive, structured, and critical overview of the field of automatic hate speech detection in natural language processing. This research found a few dedicated works that address the effect of incorporating different text transformations on the model accuracy for sentiment classification.

Aggarwal and Zhai [4] show the impact of transformation on text classification by taking into account four transformations and their all-possible combination on news and email domain to observe the classification accuracy. Their experimental analyses have shown that choosing an appropriate combination may result in significant improvement in classification accuracy. Robata & Tetreault [3] used the normalization of numbers, replacing very long unknown words and repeated punctuations with the same token.

Haddadi et. al. [5] explained the role of transformation in sentiment analyses and demonstrated with the help of SVM on the movie review database that the accuracies improve significantly with the appropriate transformation and feature selection. They have used transformation methods on raw data such as white space removal, expanding abbreviation, stemming, stop words removal, and negation handling. Other works focus more on modeling as compared to transformation.

In another study, Bojanowski et. al. [6] used five transformations namely URLs features reservation, negation transformation, repeated letters normalization, stemming, and lemmatization on Twitter data and applied linear classifier available in WEKA machine learning tool.

They found the accuracy of the classification increases when URLs features reservation, negation transformation, and repeated letters normalization are employed while decreases when stemming and lemmatization are applied.

Qian et. al. [7] investigated the effect of transformation on five different Twitter datasets to perform sentiment classification and found that removal of URLs, the removal of stop-words, and the removal of digits have less effect on accuracy whereas replacing negation and expanding acronyms can however improve the accuracy.

Most of the exploration regarding the application of the transformation has been around the sentiment classification on Twitter data which is length-restricted. The length of online comments varies and may range from a couple of words to a few paragraphs.

III. DESIGN AND METHODOLOGY

A. *Convolutional Neural Network*

Convolutional Neural Network (CNN) is an advanced and high-potential type of the classic artificial neural network model. It is built for tackling higher complexity, pre-processing, and data compilation. It takes reference from the order of arrangement of neurons in the visual cortex of a human brain.

The CNNs are considered as one of the most efficient and flexible models for specializing in image as well as non-image data. The neurons present in the convolution layers are responsible for the cluster of neurons in the previous layer.

The four stages involved in building the CNN:

1) *Convolution:* It is the process that derives feature maps from input data, and then applies a function to these maps.
2) *Max-Pooling:* It helps the CNN in detecting an image based on given modifications.
3) *Flattening:* Here, the data obtained is then flattened for the CNN to analyze.
4) *Full Connection:* It is often described as a hidden layer that compiles the loss function for a model.

B. *Natural Language Processing*

Understanding the complexities associated with a language like its syntax, semantics, expressions, or even sarcasm, is one of the hardest tasks for the machines to learn. Natural Language Processing (NLP) through Deep Learning (Deep NLP) is trying to achieve the same. Nowadays vectors representations of words, convolutional neural networks (CNN), recurrent and recurrent neural networks (RNNs), and memory augmenting strategies are helping achieve new heights in NLP. DL has now enabled machines to recognize human voices, translate languages, summarize large piece of text, and can even generate human-like text. Google's Assistant, Amazon's Alexa, Apple's Siri are some of the most popular applications of Deep NLP.

C. Methodology

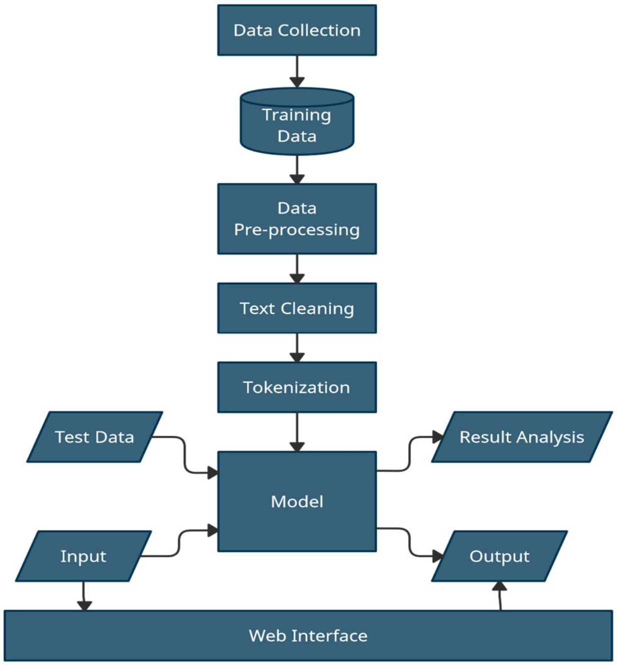


Fig. 1 Methodology flowchart

1) *Dataset Collection:* The process of gathering data depends on the type of project. For an ML project, real-time data is used. The data set can be collected from various sources such as a file, database, sensor, and other sources and some free data sets from the internet can be used. Kaggle and UCI Machine learning Repository are the repositories that are used the most for data collection for Machine learning models. Kaggle is one of the most visited websites that is used for collecting data sets.
2) *Data preprocessing:* Data pre-processing is a process of cleaning the raw data i.e., the raw data is collected and is converted to a clean data set. There are certain steps executed to convert the data into a small clean data set and make it feasible for analysis.

Most of the real-world data is messy, like:

- Missing Data
- Noisy Data
- Inconsistent Data

Some of the basic pre-processing techniques that can be used to convert raw data are:

➢ Conversion of Data
➢ Ignoring the missing values
➢ Filling the missing values
➢ Detection of outliers

3) *Text Cleaning:* The data scraped from the website is mostly in the raw text form. This data however needs to be cleaned before feeding it as an input to the model. Cleaning up the raw data is necessary to highlight the attributes that we will be wanting for the machine learning system to pick up on. Cleaning the data typically consists of several steps:

- Removing extra white spaces
- Removing punctuations
- Removing URLs and emails
- Case normalization
- Removing stop words
- Lemmatization and Stemming

4) *Tokenization:* Tokenization is one of the most common and important tasks when it comes to dealing with the text data. It is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units is called a token. The tokens could be words, numerical digits, or punctuation marks. Tokens act as the ending point of a word and the beginning of the next word.
5) *Model Selection:* Model selection is the process of selecting one final deep learning model from among a collection of candidate deep learning models for a training dataset. It is a process that can be applied across different types of models and across models of the same type configured with different model hyper-parameters.

The types of classification models are:

- Convolutional Neural Networks (CNNs)
- Long Short-Term Memory Networks (LSTMs)
- Recurrent Neural Networks (RNNs)
- Generative Adversarial Networks (GANs)
- Multilayer Perceptrons (MLPs) etc.,

6) *Train and Test Data:* For training a model we initially split the model into 2 sections which are ‘Training data’ and ‘Testing data’. The classifier is trained using the training data set, and then tests the performance of the classifier on the unseen test data set.
7) *Training set*: It is the dataset using which the computer learns how to process and analyse the information. Machine learning uses algorithms to perform the training part. Training data set is used for learning and to fit the parameters of the classifier.
8) *Test set:* A set of unseen data used only to assess the performance of a fully specified classifier.
9) *Evaluation:* Model Evaluation is an important part of the model development process. It helps to identify the best model that represents the data well and how well the chosen model will work in the future. To improve the model hyper-parameters of the model can be tuned and the accuracy can be improved. A confusion matrix can be used to improve by increasing the number of true positives and true negatives. The output is predicted by analyzing the test data as input along with test data output and then the output is displayed.
10) *Interface:* A web interface is built to take input and display an output. Flask web framework is used to build a web interface and pickle library is used to integrate both model and web page.

D. Algorithm Used

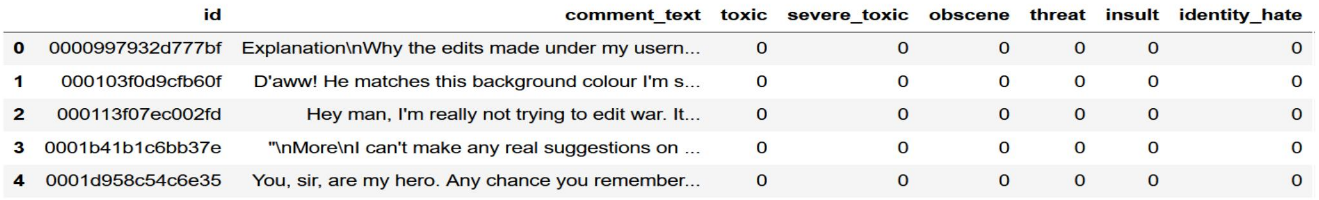
The algorithm used is Convolutional Neural Network with Natural Language Processing. CNN is used for analyzing visual imagery. For example, CNN is used for applications such as image classification, facial recognition, object detection, etc. Instead of image pixels, the input to the NLP tasks are sentences, paragraphs or documents represented as a matrix. Each row of the matrix represents one token, usually a word, but it could be a character also. That is, each row is vector that represents a word. Typically, these vectors are word embeddings (low-dimensional representations), but they could also be one-hot vectors that index the word into a vocabulary. For a 10-word sentence using a 100-dimensional embedding we would have a 10×100 matrix as our input. That’s our “image”.

IV. RESULTS AND DISCUSSION

This section gives a description of the dataset and its respective categories. The dataset includes various comments posted on the social media platform. The requirements for the model automatically detecting categories of toxicity a comment belongs are analyzed. These requirements are very essential to implementing the application.

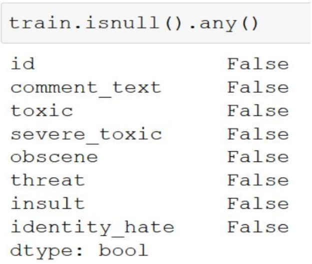
A. Implementation

The algorithms in this study were implemented in Jupyter notebook on Anaconda IDE and written in Python programming language on a personal computer with the configuration of Windows 10 Operating system and i3 processor. The results from implementing the datasets obtained in this study are discussed in this section. Figure 2 shows the distribution of the categories: toxic, severe toxic, obscene, threat, insult, and identity-hate. Figure 3 shows the checking for the missing or null values in the dataset. There were no missing values.



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

Fig. 2 Dataset



```
train.isnull().any()

id                False
comment_text      False
toxic             False
severe_toxic      False
obscene           False
threat            False
insult            False
identity_hate     False
dtype: bool
```

Fig. 3 Checking missing values in the dataset

Data visualization is performed using Python's matplotlib library. A heat map is used to find the correlation between the categories. Bar plot is used to analyze how many comments belong to each category.

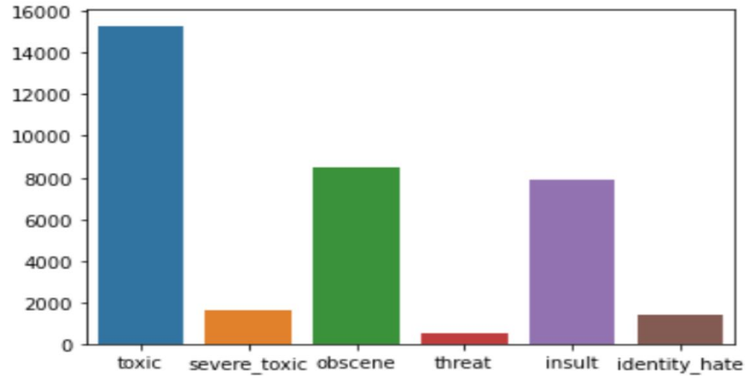


Fig. 5 Bar plot

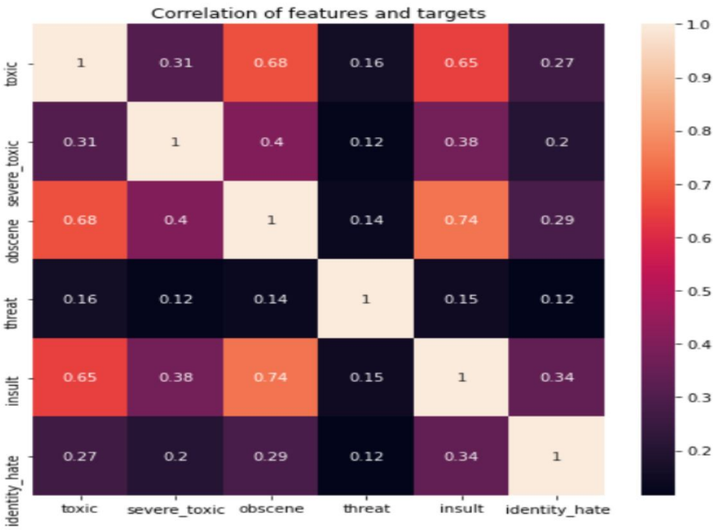
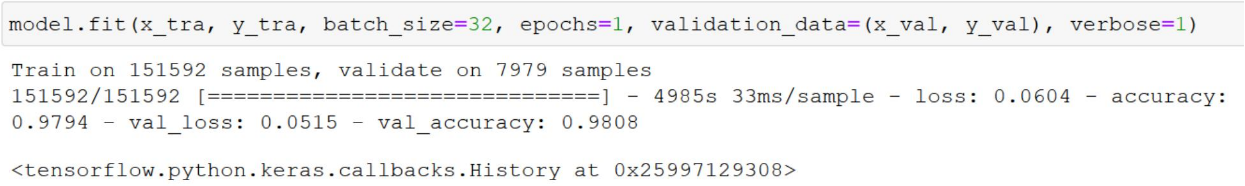


Fig. 4 Heat map

1) *Text Cleaning:* Here, nltk library provided by Python is used to remove punctuations, white spaces, and stop words from the raw comments in the dataset. URLs and emails are removed from the comments as these do not play any role in the classification. Lemmatization and stemming are also performed on the raw comments. These tasks are achieved through a user-defined function.
2) *Model Building and Training:* CNN model is built using the tensorflow. keras. model library. Training of the processed dataset is done by solving the problem as a multi-label classification problem. The pre-processed dataset is trained, and the training accuracy achieved is 98.08%. This is shown in Fig. 6.

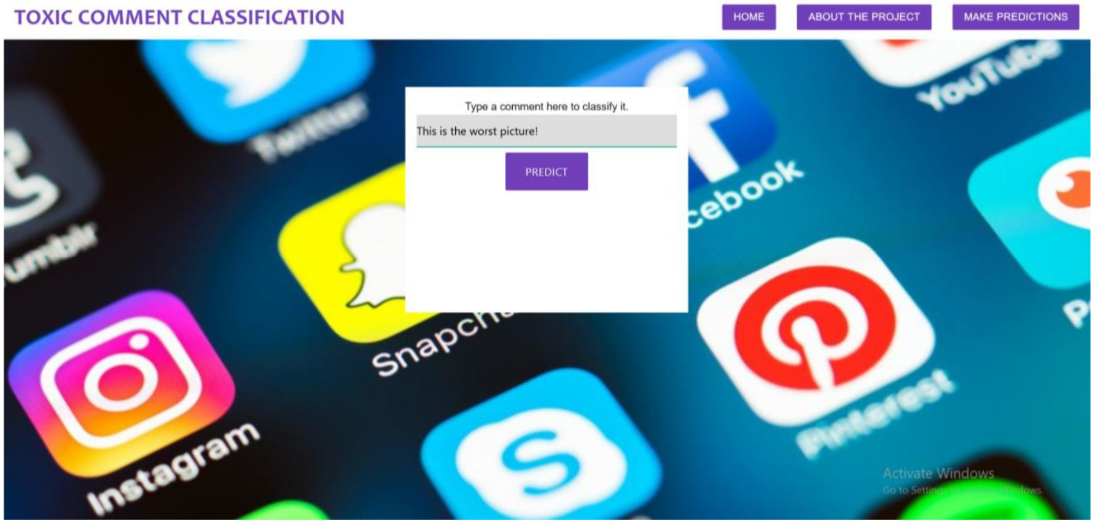


```
model.fit(x_tra, y_tra, batch_size=32, epochs=1, validation_data=(x_val, y_val), verbose=1)

Train on 151592 samples, validate on 7979 samples
151592/151592 [==============================] - 4985s 33ms/sample - loss: 0.0604 - accuracy:
0.9794 - val_loss: 0.0515 - val_accuracy: 0.9808

<tensorflow.python.keras.callbacks.History at 0x25997129308>
```

Fig. 6 Model training

B. *User Interface*

A user interface is developed using HTML, CSS, and Flask web framework. The interface is shown in the below figures.

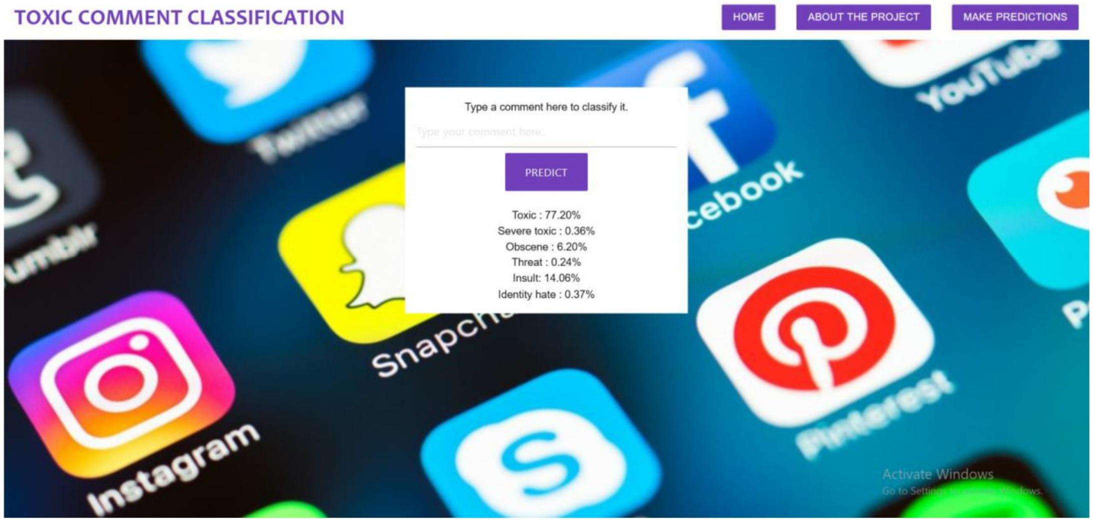


TOXIC COMMENT CLASSIFICATION

Type a comment here to classify it.

This is the worst picture!

Fig. 7 User Interface



TOXIC COMMENT CLASSIFICATION

Type a comment here to classify it.

Type your comment here

Toxic : 77.20%
Severe toxic : 0.36%
Obscene : 6.20%
Threat : 0.24%
Insult: 14.06%
Identity hate : 0.37%

Fig. 8 Predicting the toxicity

V. CONCLUSION

Communication is one of the necessities of everyone's life. People need to talk and interact with one another to express what they think. Over the years, media and social networking have been increasing exponentially due to an upsurge (rise) in the use of the internet. The responsibility lies on the social media platforms and their administrations, or the host organizations to control and monitor these comments.

This paper focuses on developing a model that would automatically classify a comment as either toxic or non-toxic using CNN with NLP. The implementation shows that the CNN algorithm performs well with an accuracy of around 98.08%. There is no definite guide on which algorithms to use in any situation. What may work on some data sets may not necessarily work on others. Therefore, we must always evaluate methods using cross-validation to get reliable estimates. In the future, this module of prediction can be integrated with a module of the automated processing system and other models for NLP. Also, discriminate analysis can be used individually or combined for enhancing reliability and accurate prediction.

REFERENCES

[1] Deng, A., Yu. D., (2014). Deep Learning. Methods and Applications. Retrieved from http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7-SIG-39.pdf

[2] Yoshua, B., (2009). Learning Deep Architectures for AI (PDF). Foundations and Trends in Machine Learning

[3] Nobata, C., Tetreault, J., Thomas, A. Mehdad, Y., Chang Y., (2016). Abusive Language Detection in Online User Content. International Conference on World Wide eb, pp. 145-153.

[4] Aggarwal,C., Zhai, C., (2012). A Survey of Text Classification Algorithms. In Mining Text Data. Springer, pp. 163-222

[5] Haddadi, C., Benevenuto, H,, Gummadi, K., (2010). Measuring user Influence in Twitter: The million-follower fallacy. 4th International AAAI Conference on Weblogs and Social Media (ICWSM), vol. 14, no. 1, p. 8.

[6] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., (2017). Enriching Word Vectors with Subword Information, TACL, vol 5, pp.135–146.

[7] Qian, M.,Sherief, E.,Belding-Royer, E., Wang. W., (2018). Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Languag Technologies, vol. 2.

[8] Tarasova, Z., Khlinovskaya-Rockhill, E. Tuprina, O., Skryabin, V., (2017). Urbanization and the Shifting of Boundaries, Contemporary Transformations in Kinship and Child Circulation among the Sakha. Europe-Asia Studies vol. 67,no.7, pp. 1106-1125.

[9] Adamic. L., (2016). The small world web, Research and Advanced Technology for Digital Libraries, pp. 852–852.

[10] Hanson, R., (2014). Foul play in information markets. George Mason University, vol. 18, no. 2, pp, pp. 107-126.

[11] Waseem, Z., Thorne, J., Bingel, J.,(2018). Bridging the gaps: Multitask Learning for Domain Transfer of Hate Speech Detection. Online Harassment, pp 29–55.

[12] Kumar, R., Ojha, A., Malmasi, S., Zampieri, M., (2018). Benchmarking Aggression Identification in Social Media, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 1-11.

[13] Georgakopoulos, V., Tasoulis, S., Vrahatis,A., Plagianakos, P., (2018). Convolutional Neural Networks for Toxic Comment Classification, ACM Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp. 35.