Multilevel Troll Classification of Twitter Data Using Machine Learning Techniques

Susan Mathew K*, Deborah Alex, Nidhi Deshpande, Richa Sharma, Arti Arya, and D. P. Balendra

Department of Computer Science and Engineering, PES University, Bengaluru, India
Email: susanmatk@gmail.com (S.M.K.); itsdeborahalex@gmail.com (D.A.); nidhideshpande15@gmail.com (N.D.); richasharma@pes.edu (R.S.); artiarya@pes.edu (A.A.); balendradp@gmail.com (D.P.B)
*Corresponding author

Manuscript received June 27, 2023; revised July 20, 2023; accepted August 11, 2023; published February 15, 2024

Abstract—**Trolling on social media is the phenomenon of using provocative or offensive text, attempts to dominate, disrupt or deviate from the main topic of discussion. Identifying trolls can help protect organic users of the platform from the unwanted negative consequences resulting from interacting with a troll. In this work, five condensed feature sets namely sentiment, readability, post analysis, network and frequency analysis are used to make the broad distinction between troll and non-troll users. An ensemble of Machine Learning Algorithms (with base classifiers as Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and meta-classifier as Random Forest) are used to perform the multilevel classification. In the first level, trolls are identified from non-trolls and in the second level, the trolls are classified into their respective types—Political, Communal, Conspiracy or Asocial Trolls. Additionally, by data driven observations, the traditional understanding of antisocial behavior in trolls is expanded to develop a more multidimensional representation of trolling behavior. Using the Stacking Classifier, an accuracy of 78.72% was achieved for identifying trolls from non-trolls in first phase and an accuracy of 83.24% in classifying trolls into their respective categories in the second phase.**

Keywords—**machine learning, trolls, types of trolls, multi-class classification, random forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), ensemble**

I. Introduction

People associate trolls with users who attack or offend others online. A troll can be a social media user who has negative or antisocial tendencies [1]. However defining and understanding trolls is far more complex since more work is being undertaken to understand the psychology behind a troll. There are sponsored trolls who spread propaganda [2–4] and then there are users who troll for self-fulfillment. Sadism is a dominant trait found in internet trolls [5–7]. Another distinctive difference is the primary subject of discussion that they troll about [8].

From manipulating individual beliefs to driving victims to sheer desperation and frustration, it has been observed that troll interactions can have a distressing effect on the physiological and psychological health of their victims. There have even been cases where victims were driven to extreme depression leading to suicidal ideation and in some cases, suicide [9]. The existence of Troll Farms [10] is further cause for concern. Troll Farms are multiple troll accounts working together in a strategic and coordinated manner usually about a singular cause.

Varied attempts have been made to classify trolls using different machine learning and deep learning approaches. Since the understanding of a troll is no longer limited to just language based indicators, these approaches range from methods based on readability and sentence structures to posting frequency and interactions.

In terms of the content trolls post and interact with, there is again further classification that can be done. Identifying these classes can be beneficial to warn users about the specific danger they may be facing. In this work, a systematic verification is done to check if the characteristics associated with trolls identified in past works still hold true today. Then a dataset consisting of fewer, broader features is created. The dataset consists of manually annotated troll users who fit the criteria laid out for each class. Details on what criteria was looked at for annotation are mentioned in Section III-A. Based on the features used in the dataset, a Machine Learning approach is used to identify trolls from non-trolls in the first phase and for classifying the trolls in the second phase. Every troll belongs to one of the classes so each class is sufficiently broad in definition and specific in application. The main contributions of this work are:

- Implementing a two-phase Stacking Classifier to identify trolls from non-trolls with a reasonable accuracy after accounting for the difference in behavior. And further classify identified trolls into their respective categories.
- Constructing a novel, compact dataset that captures the differences between the different types of trolls.
- Developing well defined troll classes—Political, Communal, Conspiracy and Asocial
- Data based observation of the different types of trolls due to varied motivations and psychological dispositions.

The remaining paper is organized as follows: Section II highlights the critical review of literature of troll detection methods. Section III gives the clear definitions of different classes of Trolls and explains the proposed approach and implementation in detail followed by Section IV that discusses the results and gives discussion of the results obtained. Section V is the conclusion and future vision of the proposed work.

II. Related Work

Tomaiuolo *et al.* [11] summarized the majors works explored until 2020 and along with their limitations. Methods are grouped under the different levels at which troll identification is possible-posts, discussion threads, user behavior and community relationships. Post based techniques have methods ranging from sentiment analysis,

DOI: 10.7763/IJCTE.2024.V16.1350

sentic computing, using Automated Readability Index (ARI), Affective Space and Hourglass of Emotions. Thread based methods combine statistical and syntactic measures like similarity and relevance. User based methods look at the user's behavior as a whole before concluding if they are a troll or not. Community level methods are the broadest level at which analysis is done which includes social network analysis.

Cambria *et al*. [12] used Sentic Computing to detect trolls. Their approach included using tools like Affective Space and The Hourglass of Emotions along with a basic Natural Language Processing (NLP) module to calculate the trollness of a user. Their formulas include Concept Frequency-Inverse Opinion Frequency (CF-IOF) weighting to detect common concepts used by the individuals and Spectral Association to expand the set of concepts obtained from CF-IOF weighting. The paper had a precision of 82% and a F-measure value of 78%. The disadvantage of this work was it seemed to focus on single posts made by users as compared to the considering the overall behavior of a particular user.

Ezzeddine *et al*. [13] transformed the Twitter user's activity into state action pairs where actions are things the account can do like tweet, retweet, interact with others or take no action. The Twitter environment is represented as a Markov Decision Process. A Long Short-Term Memory (LSTM) based classifier is used on the trajectory of actions and a troll score is computed on whose basis the identification is done. One gap in this approach is that it can be crisis specific. The dataset for training was the Internet Research Agency (IRA) trolls dataset. When the method was assessed with respect to COVID-19 suspended accounts, the AUC was 80% as opposed to the previously obtained 97%.

Cheng and Danescu-Niculescu-Mizil *et al.* [14] laid down a solid foundation and gives valuable insight into antisocial behavior. They classify users as "Future Banned Users" and "Never Banned Users" and examine these groups to understand commonly observed characteristics and how they differ between the groups. The main features that can be used to identify antisocial users were post features, activity features, community features and Moderator features. Using this information, they used a random forest classifier and logistic regression classifier to predict future banning of users. Fornacciari *et al*. [1] proposed an approach that uses six major feature groups for troll identification. This paper was instrumental in giving a clear idea of the different groups of features. The total number of features used was 224. They used Social Media Optimization (SMO), Naïve Bayes and Random Forest with a neural network as the meta learner. The neural network meta learner gave the final output with an accuracy of 93.6%. But using SMO gave an accuracy of 95.5%. This work seems to have focused on only one of the categories of trolls we have considered in our troll class definition: Asocial Trolls.

Lewinski and Hasan [15] used data from Twitter and Facebook to classify data into five classes such as Fearmonger, Gamer, Left Troll, News Feed and Right Troll. They did this with the help of models such as Latent Dirichlet Allocation (LDA), Random Forest and Support Vector Machine (SVM). The authors propose that better classification could probably be achieved by using neural networks for the classification process.

MacHova and Porezany *et al*. [16] used a dataset that consists of features such as number of characters in the post, word count, average length of words, number of capital letters in the text, number of numbers in the text, number if "I like" responses and measure of negativity, provocativeness in the comments. The algorithms used for classification are Support Vector Machine (SVM), Logistic Regression (LR) and Multinomial Naïve Bayes (MNB). Different representations such as Bag-of-Words and TF-IDF were used for the text and precision, recall and F1-rate to measure the performance of the different algorithms.

After reviewing and surveying the available literature, it is found that the majority of the available literature concentrates on one manifestation of behavior in trolls.

To the best of our knowledge, there is no work available that considers all four (Political, Communal, Conspiracy and Asocial) expressions of troll behavior. These four expressions are different in terms of specific behavioral patterns. The dataset constructed for this work is representative of different observable attributes possessed by different trolls. This paper focuses on addressing this research gap using a Stacking Classifier with Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) as the base classifiers to perform multilevel classification.

III. Proposed Approach and Implementation

A. Definitions

The classes of trolls identified are: Political, Communal, Conspiracy and Asocial.

Definition 1: Political trolls [8] post about political candidates and parties. They talk down or even malign the candidates of the opposing party or anyone against the ideology they hold to. They tend to share their internet space and band together with those that think similarly. It is observed from the literature that these users are more active with respect to retweets and replies.

Definition 2: Communal Trolls aim to divide communities on the basis of race, religion, gender etc. They depend on inciting some negative response in others like fear, anxiety, anger or rage and for this reason, their content would be less well received by the community of organic users. They are less group oriented when compared to political trolls and they also share external videos usually edited or doctored to show some community in less favorable light. Both Political and Communal Trolls are ideological trolls [8] but they differ based on whether their ideology stems from their political affiliation first. This distinction is maintained in this work.

Definition 3: It has been observed that people who advocate for conspiracy theories have a direct connection to a need for uniqueness and narcissism. They also have higher scores on the Dark Tetrad, namely Machiavellianism, narcissism, psychopathy, and sadism [17]. Conspiracies tend to thrive in the grey space. Truth is often black and white but speculation based on unconfirmed evidence is what constitutes the grey area. In this work, care is exercised to not stifle discussions or deliberations on unpopular opinions. Yet, what separates Conspiracy Trolls from the normal user debating over something controversial is this: whatever conclusion they have drawn out of inconclusive evidence is truth and they refuse to see otherwise. This particular worldview is skewed and is peculiarly dangerous and this is another dimension of behavior we want to draw attention to. These users are either paranoid because of what they believe or are condescending because they consider themselves to be the only holders of the truth. They write elaborate tweets defending their view and share the most external photos and videos in comparison to the other groups.

Definition 4: Asocial Trolls [8] are either users who do not have a single subject dominating their posts for a sustainable period of time or they are obsessive over a particular non-ideological topic and are often seen displaying behaviors of hostility and/or mockery. They do not seem to be interested in building a community. Their tweets are low effort content consisting of simple sentence structuring. They are rather careless and impulsive about the content they post. They fit the traditional understanding of antisocial behavior the most.

Inspired and encouraged by the diversity of the work already undertaken and convinced of the importance of tackling trolling in this digital age, this works sets out to use Machine Learning techniques to perform a multilevel classification. At the first level, identification of trolls from non-trolls is done. At the second level, trolls are further classified into their respective classes namely Political, Communal, Conspiracy or Asocial.

Fig. 1 outlines the entire proposed approach. As opposed to focusing on just one part of a user's activity on Twitter, for instance, the sentiment or semantics of their tweets, a well-rounded dataset that captures a broad look of the user and their activity online is key. To achieve this, five feature sets were finalized. After this, we then proceeded with the creation and annotation of the dataset. The trending hashtags in the English speaking countries were collected using the Twitter API and then users tweeting those hashtags were identified. For the annotation process, tweets, retweets, replies and media were looked at to understand if the account overall displayed troll behavior consistent with past works. If they were evaluated to be trolls, based on the criteria defined in Section III-A, a class was assigned to them. The annotated users were verified by evaluating the category assigned to them randomly. In case of conflicts in annotations, discussions ensued and the most compelling case was accepted.

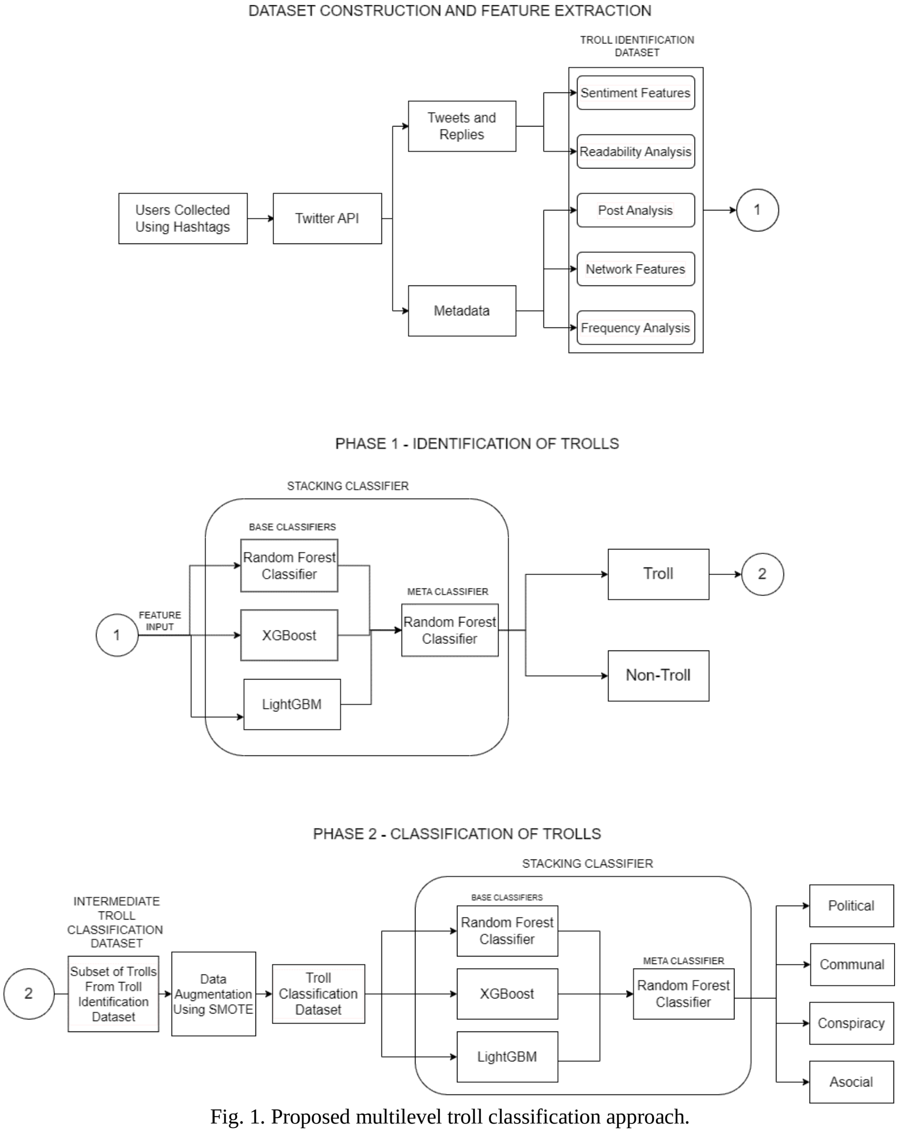

Fig. 1. Proposed multilevel troll classification approach.

B. Dataset Construction

In one such extraction of 1,121 users based on active hashtags, after annotation, 865 non-troll users and 256 troll users were observed. Amongst the 265 trolls, 82 were Political Trolls, 37 were Communal, 30 Conspiracy and 103 were Asocial Trolls. This seems representative of the organic occurrence of each category and subcategory on Twitter.

To give equal representation and to better understand each class well, controversial hashtags relevant to each category of trolls were manually looked under to find more users. After this, the dataset finally had 767 trolls and 865 non-trolls. This user dataset had 202 Asocial Trolls, 192 Political Trolls, 187 Communal Trolls and 186 Conspiracy Trolls and 1,632 users in total. The dataset used in this work was collected over a period of 6 months from June, 2022 to December, 2022.

After the user list was ready, using Tweepy, the text and metadata was extracted to compute features belonging to the 5 feature sets. The Troll Identification Dataset consists of 5 broad feature sets—Sentiment Features, Readability Analysis, Post Analysis, Network Features and Frequency Analysis. The relevant numeric measures corresponding to each feature in every set were computed and added [1].

C. Feature Extraction

1) Sentiment Features: Past works concluded that sentiment analysis alone is not sufficient for troll identification but when combined with other features or methods, it can be useful. For the dataset, 20 tweets were extracted. Computing the sentiment scores for just one tweet may not be representative of the particular user. Keeping in mind that different users have different activity rates, averaging the sentiment over twenty tweets gives more consistent measures for sentiment. VADER [18] was used for sentiment analysis because it works extremely well with social media text. The positive, negative, neutral and compound sentiment for each tweet was computed and the values were then averaged over 20 tweets to get the sentiment measures for that user.

2) Readability Analysis: In this collection, the focus is on 27 indexes that revolve around tweet readability and text analysis. They were computed using python libraries, namely py readability and textstat. Here, the number of syllables, number of words used, the number of sentences used were looked at and common readability measures like Flesch, Cl, Dale, ARI [1] etc. were applied to identify the textual pattern of the individuals tweets.

a) Flesch: It is a readability index that helps us learn how difficult a passage in English is to understand. Word length was used to determine how difficult a word is [19]. The formula it follows is:

$$FI = 206.835 - 1.015 \times \left(\frac{t}{s}\right) - 84.6 \times \left(\frac{l}{t}\right) \quad (1)$$

where, t: total words; s: total sentences; l: total syllables.

b) Dale: This is another readability index that gives us the difficulty of comprehension of a given passage. In this case, the researchers compiled a list of 3,000 commonly used words and any word not in the list was deemed as difficult [19].

$$DI = 0.1579 \times \left(\frac{d}{w} \times 100\right) + 0.0496 \times \left(\frac{w}{s}\right) \quad (2)$$

where, d: difficult words; w: words; s: sentences.

c) ARI: It is readability index that helps us determine the understandability of the text by also taking into consideration the actual number of characters in a word as compared to most other indices that only use syllables [19].

$$ARI = 4.71 \times \left(\frac{c}{w}\right) + 0.5 \times \left(\frac{w}{s}\right) - 21.43 \quad (3)$$

where, c: characters; w: words; s: sentences.

d) Cl: The Coleman–Liau index is another readability index that is similar to ARI. It uses characters to come up with a score that helps us determine how easy a text is to understand [20].

$$Cl = 0.0588 \times l - eq0.296 \times s \quad (4)$$

where, l: average number of letters per 100 words; s: average number of sentences per 100 words.

3) Post Analysis: This set consists of five features which break down the content of the user's post such as the number of photos, videos, gifs, statuses count and count. Statuses count is the number of tweets and retweets the user has. Count tells us the amount of content produced by the user.

4) Network Features: This feature set contains the relevant values that quantifies a user's behavior and interactions with the community. In this feature set there are 9 features that include followers, following, favorite count, ratio of following to followers etc. Followers are the number of accounts the user is following. Following is the number of users who are following the user being analyzed. Favorite count has the number of tweets the user has marked as favorite. The frequency of hashtags and user mentions are also calculated.

5) Frequency Analysis: In this set, there are 18 features that capture the frequency of user activities throughout the day. A 24-hour period is divided into 6 four-hour time slots. For each of these time slots the average number of tweets, retweets and replies are determined.

D. Phase 1: Identification of Trolls

These 5 feature sets along with whether the user is a troll or non-troll (denoted by 1 and 0) makes up the Troll Identification dataset. This dataset is given as feature input to the Machine Learning model. The Machine Learning model is a Stacking Classifier with Random Forest Classifier (RFC), LightGBM (LGBM), and XGBoost (XGB) as the base classifiers with Random Forest (RFC) as the meta-classifier. Using a Stacking Classifier optimally combines the predictions of the base classifiers to give the final output. All the base classifiers (RFC, LGBM, XGB) were optimized with hyperparameter tuning using Random Search. Different classification algorithms were used as the meta-classifier and Random Forest is experimentally found to be the best performing meta-classifier. The model performance is evaluated using both accuracy and F1 score. The F1 score takes precision and recall into account. The Stacking Classifier gave an accuracy of 78.72% after cross validation (5 folds) in identifying if the user was a troll or non-troll. Table 1 contains the detailed tabulation of accuracy and F1 score. All the results reported are those obtained after cross validation (5 folds). Although Random Forest alone seems to give a higher accuracy, but F1 score considers both false positives and false negatives as opposed to just the true positive and true negative values. Going by this parameter, the Stacking Classifier is the better classifier.

Table 1. Accuracy and F1 for identification of trolls from non-trolls

Classifiers	Accuracy (%)	F1
Random Forest Classifier (RFC)	78.90	0.7798
XGBoost (XGB)	77.57	0.7766
LightGBM (LGBM)	78.15	0.7794
Stacking Classifier (RFC, XGB, LGBM)	78.72	0.7872

E. Phase 2: Classification of Trolls

In this next phase, the trolls will further undergo classification into one of the four types—Political, Communal, Conspiracy or Asocial. All the users flagged to be trolls along with their feature sets from the Troll Identification Dataset will form the Intermediate Troll Classification Dataset. The Inter-mediate Troll Classification Dataset is augmented category-wise using (SMOTE) with a custom sampling strategy. Before data augmentation, the dataset had 202 Asocial Trolls, 192 Political Trolls, 187 Communal Trolls and 186 Conspiracy Trolls. After data augmentation with SMOTE, each category has 500 trolls and the dataset size increased to 2000 troll users in total. This dataset is the Troll Classification Dataset. The Troll Classification Dataset is given to the Stacking Classifier with Random Forest, XGBoost, and LightGBM as base classifiers with Random Forest as the meta-classifier to perform multiclass classification. The model performance is evaluated using accuracy and micro F1. Micro F1 is used for multiclass classification with a balanced dataset. The Stacking Classifier yielded an accuracy of 83.24%. The detailed accuracy and micro F1 tabulations can be found in Table 2. Although just using Random Forest gives a better accuracy, using the Stacking Classifier gives a higher F1 score.

Table 2. Evaluation metrics for multiclass classification

Classifiers	Accuracy (%)	F1 (micro)
Random Forest Classifier (RFC)	84.59	0.8385
XGBoost (XBG)	79.69	0.7969
LightGBM (LGBM)	84.25	0.8425
Stacking Classifier (RFC, XGB, LGBM)	83.249	0.8400

IV. Results and Discussion

When comparing data between trolls and non-trolls by using the arithmetic mean of the features in Table 3, we see that troll users have a lower positive sentiment and higher negative sentiment than non-trolls. The values recorded in Table 3 have been truncated to two decimal points with the exception of followers and following which have been truncated to the corresponding whole number. It is observed that non-troll users have more retweets than troll users throughout the day. Trolls have more replies than non-trolls. In Table 3, it is seen that troll users have a higher frequency of user mentions in their tweets compared to non-troll users. One of the contributing factors to this is that trolls tend to personally mention and attack others more than non-trolls. Troll users use more hashtags than non-trolls which was expected since it is theorized that some classes of trolls were driven by a desire for greater visibility. Readability indices like Dale, ARI, and Cl, use a scale of increasing values to indicate increase in difficulty of comprehension of text. On the other hand, for the Flesch readability index, it is vice versa where lower scores indicates more complex language and higher scores more easy to understand language. It is observed that trolls write less readable and complex sentences as indicated by the different readability indexes (ARI, flesch, dale, cl) in Table 3. The results are on par with those recorded in [1, 14]. The broad behavioral manifestations of trolls in the online space have continued to remain consistent.

Table 3. Comparison of major features between trolls and non-trolls

Feature	Troll	Non-Troll
Following	2295.38	1573.71
Followers	25287.00	26411.10
Frequency of Hashtags	0.32	0.20
Frequency of User Mentions	1.13	0.78
Favorite Count	28886.73	51690.85
Status Count	212289.29	49297.30
Positive Sentiment	0.13407	0.154211
Negative Sentiment	0.10611	0.09252
Videos	0.42	0.24
Photos	3.75	2.22
ARI	19.41	16.52
flesch	40.79	49.72
dale	16.48	15.42
cl	17.88	15.43

In this work, the traditional understanding of troll behavior is expanding to capture the multidimensional way this behavior manifests. After a comparison of the mean obtained by each troll type for different features, it is possible to conclude that each class of trolls has a unique pattern of behavior. Table 4 contains the class wise comparison of the mean of some major features where POL refers to Political Trolls, COM refers to Communal Trolls, CON refers to Conspiracy Trolls and AS refers to Asocial. All the values have been truncated to 2 decimal points with the exception of Positive Sentiment, Negative Sentiment and Videos which are truncated to include 4 decimal places.

Political Trolls follow the highest number of users and they have the second highest number of followers. They also favorite more tweets than other classes and they have more tweets and retweets (high status count). Studying the frequency analysis of their activity, an obvious pattern emerges where Political Trolls consistently have the highest retweets and replies activity. They have the second highest number of user mentions (User Frequency) which validates the high reply activity. Status count is the number of tweets and retweets of the user. Seeing that Political Trolls have the highest retweet activity, it can be concluded that instead of posting more original content, Political Trolls amplify voices and people they agree with by retweeting.

Table 4. Comparison of major features between the different types of trolls

Feature	POL	COM	CON	AS
Following	3527.04	2257.53	2004.62	1427.46
Followers	37737.58	14310.05	9463.62	38184.66
Favorite Count	38517.88	29180.37	24086.97	23880.10
Status Count	28774.42	20342.65	21583.17	14780.46
Hashtag Frequency	0.17	0.38	0.50	0.24
User Frequency	1.22	0.99	0.90	1.39
Positive Sentiment	0.1285	0.1321	0.1262	0.1517
Negative Sentiment	0.1182	0.0976	0.0883	0.1188
Videos	0.1718	0.4759	0.9193	0.1732
Photos	3.63	3.51	5.10	2.83
ARI	18.55	18.73	22.11	18.38
dale	15.74	16.68	18.28	15.35

Communal Trolls follow the second highest number of users but they only rank third when you consider the number of followers they have. While Political Trolls tend to exist in groups, Communal Trolls do not. They post the second highest number of videos and are similar to Political Trolls with respect to the number of photos posted. Communal Trolls rely on external videographic material to incite division. Organic users of social media are not receptive towards videos content that causes negative emotions. This could explain their low follower count. Additionally, they have the second highest number of favorites and the second highest hashtag use. They could be aiming to increase visibility of their posts and ideology. They consistently have large periods of least activities when we observe the frequency analysis, which could be linked to geographic division. Lastly, they tend to write less readable content than Political Trolls.

Conspiracy Trolls are following the third highest number of users and have the lowest number of followers. This could be an indication that the content they post is unpopular with other users of the platform. It is interesting to note that they post the highest number of photos and videos. They rely heavily on external media to lend some strength to their view. They use the highest number of hashtags to give their content more visibility. They have the lowest user mentions. Based on ARI, Conspiracy Trolls write sentences that are the most difficult to read. Based on dale, we can confirm that Conspiracy Trolls also tend to use more uncommon words in their sentences. Lastly, under frequency analysis, Conspiracy Trolls have the highest tweet activity.

Asocial Trolls follow the lowest number of users on average but they have the highest number of followers. They share the lowest number of photos and the third lowest number of videos. Additionally, they have the lowest favorite count which could further indicate shallow interests. They have the lowest status count which means they have the lowest number of tweets and retweets. Since they have the highest user frequency, most of their activity has to be replies. An interesting observation is that Asocial Trolls rank highest in terms of both most negative and most positive content posted. This lends credibility to the diverse nature of topics and to how they tend to ascribe to extreme generalizations. Amongst all the 4 categories, most of the readability indexes conclude that Asocial Trolls write easier to read, simpler sentences.

For identifying trolls from non-trolls, as tabulated in Table 1, Random Forest gave an accuracy of 78.90%, XGBoost gave an accuracy of 77.57%, and LightGBM gave an accuracy of 78.15%. The Stacking Classifier with Random Forest as the meta classifier and LightGBM, XGBoost, and Random Forest as the base classifiers gave an overall accuracy of 78.72% with an F1 score of 0.7872.

A subset of the Troll Identification Dataset was taken to include the 202 Asocial trolls and 865 non-troll users. The data imbalance was addressed by oversampling using SMOTE. An interesting observation was when this dataset was passed to the Stacking Classifier, an accuracy of 91.73% was obtained after cross validation (5 folds). All other classes were similarly given to the Stacking Classifier after balancing with SMOTE. The model identified Political Trolls from non-trolls with an accuracy of 92.77%, Communal Trolls with an accuracy of 92.71%, and Conspiracy Trolls with an accuracy of 93.58%. The detailed tabulation of evaluation metrics is available in Table 5. The results obtained considering each troll category separately are on par with the results obtained by previous works.

Table 5. Class wise metrics for identification on of trolls

Classifiers	Political		Communal		Conspiracy		Asocial	
	Accuracy (%)	F1	Accuracy (%)	F1	Accuracy (%)	F1	Accuracy (%)	F1
RFC	91.79	0.9170	92.13	0.9209	94.33	0.9435	91.50	0.9136
XGB	91.84	0.9146	92.31	0.9165	92.36	0.9206	90.17	0.8972
LGBM	92.83	0.9266	92.48	0.9164	93.23	0.9301	91.90	0.9174
Stacking Classifier	92.77	0.9255	92.71	0.9232	93.58	0.9347	91.73	0.9135

Table 6. Comparative analysis of past works with our approach

	Work	**Fornacciari *et al.*** [1]	**MacHova *et al.*** [16]	**Our Approach**
Troll Type	Political		*	✓
	Communal		*	✓
	Conspiracy		*	✓
	Asocial	✓	*	✓
Dataset Size	No of trolls	500	**	767
	No of non-trolls	500	**	865
	No of features	224	7	63
Best Classifiers		SMO	Multinomial Naive Bayes using Bag of Words and TF-IDF representation	Random Forest Classifier
Evaluation Metrics		Acc: 95.5	Bag of Words—Recall:0.92 Precision:0.63 TF-IDF Recall:0.92 Precision:0.60	Accuracy: 78.90 F1: 0.7798

*- The dataset consists of comments related to SARS-Cov2 coronavirus pandemic in Slovakia. It is unclear which troll types have been considered for this dataset.

**- 2500 comments were collected and filtered to obtain a balanced dataset. No details have been provided about the exact split.

In view of the results obtained for identifying individual troll types from non-trolls as recorded in Table 5, a comparative analysis has been recorded in Table 6. The overall accuracy of 78.72% must be interpreted in light of the inclusion of all four types of trolls in the dataset used in this work.

Even though Random Forest gives a better accuracy for phase II when the overall dataset is considered, should our proposed approach be used for identifying just one type of trolls from non-trolls, the Stacking Classifier will work better as can be observed in Table 5. All the results reported here are cross validated scores with the number of folds as 5.

The subset of the Troll Identification Dataset which contained only trolls was augmented so as to get 500 trolls per class. With the same Stacking Classifier combination that was used to identify if a user was a troll or non-troll, an accuracy of 83.24% was obtained in classifying trolls into their respective class—Political, Communal, Conspiracy or Asocial.

V. Conclusion

In this work, the characteristics of troll behavior and their differences from non-trolls users with respect to past works were verified. The broad characteristics of troll behavior have remained consistent. Four types of trolls were identified—Political, Communal, Conspiracy, Asocial and definitions for each type were laid out. A dataset of five feature sets with 63 features in total was constructed. A detailed analysis of the data grouped according to the type of troll shed light on previously undocumented, observable behavioral characteristics of each type of troll. A multidimensional representation of trolls was successfully captured in the dataset used in this work. The proposed approach of using a Stacking Classifier with the base classifiers as Random Forest, XGBoost, and LightGBM with Random Forest as the meta-classifier to identify troll users from non-trolls gave an overall accuracy of 78.72%. Additionally, classification of trolls into their respective type using the Stacking Classifier with the base classifiers as Random Forest, XGBoost, and LightGBM, and Random Forest as the meta-classifier gave an accuracy of 83.24% and a micro F1 score of 0.84.

At the end of this work, one derived conclusion is that Troll Identification and Classification is a topic that requires more diverse and extensive data than what is publicly available presently. Additionally, future work in this field must consider the different manifestations of troll behavior and researchers must exercise careful caution in ensuring the same.

Further work can be done to classify trolls into their respective troll types based on the users post content. As an extension of this work, attempts were made to use Deep Learning on the users tweets for this. However, this task demands more data than what has been acquired for this work.

Throughout the course of this work, there existed an overlap between Political and Communal Trolls in terms of the content they were posting. The occurrence of a significant, influential external event of importance, like an election, would cause a shift in the content of Communal Trolls to resemble that of Political Trolls. This intersection of behavior can be studied and explored with the intent of quantifying it.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Susan Mathew K contributed towards feature set construction and data annotation. She performed data analysis and implemented the machine learning techniques. Alongside interpreting the results, she contributed to the writing of the research paper. Deborah Alex helped in the collection and annotation of data and in the implementation and consolidation of the results obtained using machine learning techniques. She also contributed to the writing of the research paper. Nidhi Deshpande helped in the extraction, annotation and aggregation of the data set. She contributed to the implementation and consolidation of the results and towards writing of the research paper. Richa Sharma and Arti Arya contributed towards the writing of the research paper and the architectural design while operating in the capacity of guides for this work. Balendra DP helped in collection and annotation of the data set. All authors had approved the final version.

References

[1] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on Twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018. doi: 10.1016/j.chb.2018.08.008

[2] S. Zannettou, M. Sirivianos, T. Caulfield, G. Stringhini, E. de Cristofaro, and J. Blackburn, "Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web," in *Proc. the Web Conference 2019—Companion of the World Wide Web Conference*, 2019, pp. 218–226. doi: 10.1145/3308560.3316495

[3] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out? towards understanding state-sponsored trolls," in *Proc. WebSci 2019—the 11th ACM Conference on Web Science*, 2019, pp. 353–362. doi: 10.1145/32925p22.3326016

[4] C. Llewellyn, L. Cram, R. L. Hill, and A. Favero, "For Whom the bell trolls: Shifting troll behaviour in the Twitter Brexit debate," *Journal of Common Market Studies*, vol. 57, no. 5, pp. 1148–1164, 2019. doi: 10.1111/jcms.12882

[5] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, "Trolls just want to have fun," *Personality and Individual Differences*, vol. 67, pp. 97–102, 2014. doi: 10.1016/j.paid.2014.01.016

[6] E. E. Buckels, P. D. Trapnell, T. Andjelovic, and D. L. Paulhus, "Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment," *Journal of Personality*, vol. 87, no. 2, pp. 328–340, 2019. doi: 10.1111/jopy.12393

[7] N. Sest and E. March, "Constructing the cyber-troll: Psychopathy, sadism, and empathy," *Personality and Individual Differences*, vol. 119, pp. 69–72, 2017. doi: 10.1016/j.paid.2017.06.038

[8] M. R. Sanfilippo, S. Yang, and P. Fichman, "Managing online trolling: From deviant to social and political trolls," in *Proc. Hawaii International Conference on System Sciences (HICSS)*, 2017.

[9] C. L. Gomez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Logic Journal of IGPL*, vol. 239, pp. 419–428, 2016. doi: 10.1007/978-3-319-01854-6_43

[10] S. McCombie, A. J. Uhlmann, and S. Morrison, "The US 2016 presidential election & Russia's troll farms," *Intelligence and National Security*, vol. 35, no. 1, pp. 95–114, 2020. doi: 10.1080/02684527.2019.1673940

[11] M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, and Poggi, "A survey on troll detection," *Future Internet*, vol. 12, no. 2, 2020. doi: 10.3390/fi12020031

[12] E. Cambria, P. Chandra, A. Sharma, and A. Hussain. Do Not Feel the Trolls. [Online]. Available: http://cs.stir.ac.uk/eca/sentics

[13] F. Ezzeddine, L. Luceri, O. Ayoub, I. Sbeity, G. Nogara, E. Ferrara, and S. Giordano, "How 'troll' are you? Measuring and detecting troll behavior in online social networks," arXiv preprint, arXiv: 2210.08786, 2022. http://arxiv.org/abs/2210.08786

[14] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proc. the Ninth International AAAI Conference on Web and Social Media*, 2015, pp. 61–70.

[15] D. Lewinski and R. Hasan, "Russian troll account classification with Twitter and Facebook data," arXiv preprint, arXiv:2101.05983, 2021. https://arxiv.org/abs/2101.05983

[16] K. MacHova, M. Porezany, and M. Hreskova, "Algorithms of machine learning in recognition of trolls in online space," in *Proc. SAMI 2021—IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 2021, pp. 349–353. doi: 10.1109/SAMI50585.2021.9378699

[17] L. Pummerer, "Belief in conspiracy theories and non-normative behavior," in *Current Opinion in Psychology*, 2022. doi: 10.1016/j.copsyc.2022.101394

[18] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. the International AAAI Conference on Web and Social Media*, 2014, pp. 216–225. doi: 10.1609/icwsm.v8i1.14550

[19] G. R. Klare, "Assessing readability," *Reading Research Quarterly*, vol. 10, no. 1, pp. 62–102, 1974. doi: 10.2307/747086

[20] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, vol. 60, no. 2, pp. 283–284, 1975. doi: 10.1037/h0076540