2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

# Looking Behind the Mask: A framework for Detecting Character Assassination via Troll Comments on Social media using Psycholinguistic Tools

Ahmed Al Marouf
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
Email: marouf.cse@diu.edu.bd

Rasif Ajwad
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
Email: ajwad.raw@gmail.com

Adnan Ferdous Ashrafi
Department of Computer Science and Engineering
Stamford University Bangladesh
Dhaka, Bangladesh
Email: adnan@stamforduniversity.edu.bd

***Abstract*—With the facilities of social media platforms like Facebook, Twitter, Google+, YouTube etc. people are capable of expressing their views & news, sharing moments via photos, liking, commenting and sharing others posts. The online social networks (OSNs) are not only giving positive supports to its users, but also creating opportunities to assassin personals by the trolls. Trolls are usually the OSN users who try to hide themselves while doing bad comments, false accusations, starting controversies, spreading fake news or rumors which could be considered as character assassination of individuals. The online behavior of an OSN user could be tracked via his/her digital footprints. Though tracking huge number of users who are generating billions of textual and image data every day, could be considered as a challenging task. In this paper, we have proposed a novel detection system for identifying character assassination from social media platforms. The proposed method first predicts the personality traits using users' textual data. Therefore, LIWC, SlangNet, SentiWordNet, SentiStrength, Colloquial WordNet has been utilized as psycholinguistic tool. LIWC-based feature engineering has been performed on the comments of the trolls as well as the victim user. SlangNet and Colloquial WordNet is used for detecting English slang words in the comments as it is evident that slangs are the basic communicative way to defame someone.**

***Keywords—character assassination; social media; digital footprints; psycholinguistic tools, LIWC, SlangNet, SentiWordNet, SentiStrength, Colloquial WordNet***

## I. INTRODUCTION

With the rapid growth of social media in the current era of modern technology, people are rigorously using the online platforms for communications. Online social networks (OSNs) are not only used for communication or social connection, but also nowadays media personals or politicians are creating fan-base to keep up the reputation. Some common facilities provided by the OSNs are posting current news or views of a social entity, commenting, liking or sharing other posts, sharing moments via photos, sharing current location via check-in etc. Because of these facilities and easy access provided by the OSNs such as Facebook, Twitter, Google+, YouTube, the popularity of social media is increasing day by day. Though the positive sides of the OSNs are attracting the mass, hence trolls are also interested in violating the basic ethical social behavior in OSNs. Because of the easy access and opportunity to hide oneself, trolls are becoming very regulatory in defaming others by hiding behind the mask of fake profiles. The procedure of creating fake profile is same as the regular profiles but using others data or incomplete data. For example, one can create a social network profile without using his/her own photo or valid name, address. Trolls are utilizing this data breach and connecting with socially acceptable popular personals such as media actors, politicians or even hyped online celebrities.

Trolls are usually the OSN users who utilizes the facilities of OSNs and posts bad comments, accuses someone online with false information, starts rumors or controversies, spreads fake news or influencing people to viral posts etc. Each of the above mentioned actions could be considered as character assassination. It is common scenario that actresses of renowned media background are accused falsely. The celebrities sometimes try to block or ban those commenters manually or ignore them. Commenting or sharing false news about someone could be considered as direct character assassination which is punishable crime. In this paper, we have tried to detect the trolls who are performing such despicable things over OSNs.

Despite having huge number of users connected in social networks, the authority of the OSNs are maintaining graph based architectural structure behind the social media. As a result, tracking any users' activity has become easier. The activities performed by any user could be traced as digital footprints. Hence, the online behavior of a user could be depicted from his/her activities performed online [1]. The user generated data such as textual data (status updates, comments, shared posts) and image data (photos) could be used to predict the behavior of a user. It is evident in literature that the textual data could be efficient to predict the personality of users' from social media [2, 3]. Psychological involvement is a big factor while using social media, as users keep in mind about their online reputation.

**978-1-5386-9111-3/19/$31.00 ©2019 IEEE**

Psycholinguistic tools such as linguistic inquiry and word count (*LIWC*) [4] and word corpuses such as *WordNet* [5], *SlangNet* [6], *The Colloquial WordNet* [7] could be efficiently used in the context of detecting the personality traits of social media users. Using the tools we can foster the features required to train a classification model and test it accordingly. Using the theory of supervised learning, we may get to a convenient situation to predict the troll comments from huge comment thread.

In this paper, we have proposed a novel framework for identifying the troll behind the mask, therefore, detecting troll comments from public posts. By detecting the troll comments, we may recommend a list of commenters who must be banned. We have utilized the psycholinguistic tools such as *LIWC* and *SlangNet* for extracting psycholinguistic features and slang words, respectively. The rest of the paper contains section II discussing the related works, section III describing the problem definition of character assassination in social media. Section IV presents the proposed framework and the empirical analysis are illustrated in section V. Finally section VI concludes with summary.

## II. Related Works

In this section, we have discussed on the related works been found in literature. This section is divided into two parts: literature of personality prediction and literature of Psycholinguistic tools.

Predicting personality traits from social media is a challenging task and many computer scientists and psychology researchers' have attempted to devise different methodologies. Personality trait prediction could be performed manually using a set of questionnaire. The test takers need to answer a set of questions honestly to predict his/her personality properly using scoring based methods. International Personality Item Pool (IPIP)[1] are the items or questions to be answer to devise a scoring mechanism for traits identification. Depending on the behavior of test taker on different issues of practical life, these items are presented. Ground truth datasets are proposed by researchers for computational prediction tasks. P. Howlader et al. [2] proposed a linear dirichlet allocation (LDA) based topic modeling approach on Facebook status updates and applied flexible regression models for prediction. They used machine learning techniques such as support vector regression, decision tree, LDA, LIWC, TF-IDF (Term Frequency- Inverse Document Frequency) to train the system. T. Tandera et al. [3] proposed a method to predict personality traits applying deep learning algorithms. The different feature engineering techniques are used for prediction by different researchers. The emerging trends of personality prediction from online social networks are reviewed by V. Kaushal and M. Patwardhan in [8]. They listed different categories of features such as linguistic features (LIWC features, POS tags, Speech acts, sentiment features), non-linguistic features (structural, behavior, temporal features) and social network features. Therefore, literature have sufficient evidential proof of computational personality prediction.

[1]https://ipip.ori.org/

Psycholinguistic tools are the software tools developed for easy experimentations of popular research methods. LIWC [4] is one of the psycholinguistic tool which consists a psycholinguistic dictionary in backend which contains huge number of words, synonyms and antonyms in different psychological categories. LIWC is proven to be useful in the context of personality traits prediction. The Big Five Factor Model (BFFM) consist five personality traits namely, extraversion, conscientiousness, openness to experience, agreeableness and neuroticism.

Psycholinguistic words dictionary could be useful to identify these five different traits using the related words list. WordNet [5] is a lexical dictionary of words having large number of words, which could be used for understanding the sentence patterns and language style of the status updates. Similar word list is being proposed by S. Dhuliawala et al. [6] named SlangNet. SlangNet is a WordNet like architecture containing numerous English slang words. Identifying slang words in comments could be interesting finding, as it has higher probability to be detected as troll comment. The colloquial WordNet [8] is another lexical dictionary having huge number of colloquial words. Identifying these colloquial words may tends to identifying negative comments. Identifying troll comment could be considered as one of the application context of SlangNet and Colloquial WordNet.

Sentiment features could be useful parameter to detect troll comments. In the literature, there are many methods and tools implemented to identify the inner meaning, thus sentiment of written text. In this paper, we have used SentiWordNet [9, 10] for identifying the sentiment values of the posts as well as comments. Though in the literature there are many profound sentiment analysis methods, a very little works has been done up on sentiment detection from social media informal text. SentiStrength [11, 12] is tool which can automatically analyze 16,000 social web tests per second and has human-level accuracy achieved for short social web texts in English, except for political contexts.

In this paper, we have proposed a novel approach to detect the troll comments posted for character assassination utilizing the above mentioned psycholinguistic tools and supervised learning based predictive model. One of the contribution of this paper is to utilize the existing tools for the problem. The problem definition is described in the next section.

## III. Character Assassination in Social Media

Character assassination in social media could be defined as "blaming someone with false accusation for defaming using comments or posts". Detection of character assassination could be defined as "detecting the troll comments which are responsible for damaging someone's reputation in social media". The troll could be someone who does not even know the victim personally, but being judgmental seeing the posts of victim, he/she gets interested to start trolling. In this paper, we are going to use "troll" to mention the person who is responsible for creating bad comments, whereas "victim" to mention the person whose online social image or reputation is attacked. Let us illustrate a scenario of character assassination in social media.

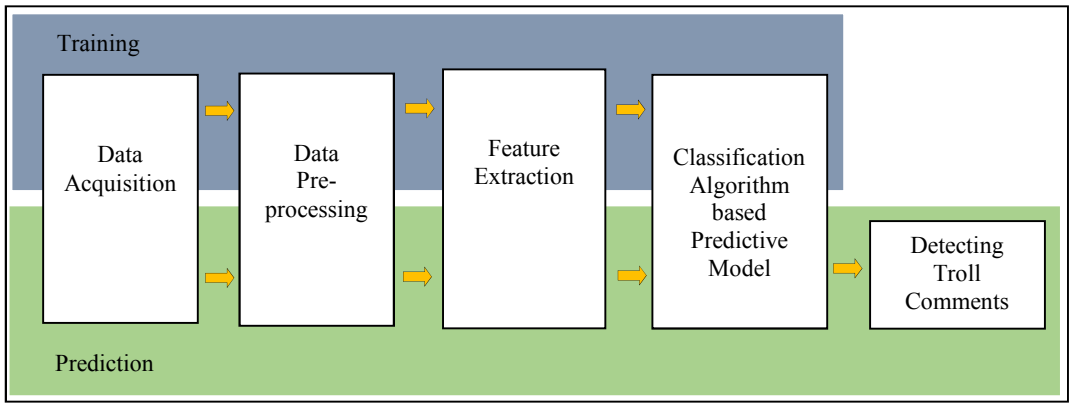



Fig.1. Proposed approach for detecting troll comments.

An actress of Hollywood (X), has posted some of her personal photos of the party she attended last night. In the photos she is found with her close friends hanging out. As she is very popular because of her successful acting career, she got millions of fans on her Facebook Page or Instagram. A troll (T), who is also a fan of X (liked her page), watching the photos become so judgmental that he/she commented using some slang words. Detecting character assassination from social media textual data could be a promising task and personalized recommender systems could be developed to prevent vocal harassments in public platforms such as social media.

## IV. Proposed Approach for Troll Comment Detection

In this paper, we have proposed a novel approach for detecting character assassination in social media through identifying the troll comments. The proposed framework depends on detecting troll comments first. Therefore, the troll comment detection approach is based on supervised learning approach based predictive model. This approach could be divided into training and testing models. The framework consist five steps namely data acquisition, data preprocessing, feature extraction, classification algorithm based predictive model.

### A. Data Acquisition

Collecting social media data is itself a research problem. There are some possible solution such as using web crawlers, application programming interface (API) etc. Graph API is used as backend of Facebook and using this API we can extract status, likes and comments of users which are only publicly available. For comments we have used *BeautifulSoap*[2] as for crawling the HTML tags. The number of comments extracted is around 7000 of 140 commenters.

### B. Data Preprocessing

The unwanted tags such as promotional posts, website links, and Uniform Resource Locator (URL) links are removed from the comments from the datasets. All the comments are in English and converted into lowercase letters.

[2]https:// pypi.org/project/beautifulsoup4/

TABLE I. Properties of Comment Dataset

| Properties | Values |
|---|---|
| No. of Comments | 6985 |
| Total Words | 62, 968 |
| Avg. Words (per comment) | 9.0147458 |

### C. Feature Extraction

In this step, we have extracted different types of features from the comments. The features are of six types: LIWC features, Parts-of-speech tags, SlangNet percentage, Colloquial WordNet percentage, SentiWordNet percentage and SentiStrength features (positive value and negative value). In total 19 features are extracted from each of the comments and the class value (yes, no) are determined to devise a prediction model. The details about the features are listed in Table II.

TABLE II. Catefories of Features to be Extracted

| Categories | Features |
|---|---|
| LIWC Features ($f_1, \ldots f_6$) | Emotion Affect, Positive emotion, Negative emotion, Anxiety, Sad, Anger |
| POS Tags ($f_7 \ldots, f_{14}$) | Percentage of Noun, Pronoun, Adjective, Verb, Adverb, Preposition, Conjunction, Interjection |
| SlangNet ($f_{15}$) | Percentage of English Slang words |
| Colloquial WordNet ($f_{16}$) | Percentage of English Colloquial words |
| SentiWordNet ($f_{17}$) | Percentage of English Sentiment related words |
| SentiStrength ($f_{18}, f_{19}$) | Positive Strength Value, Negative Strength value |

Though LIWC outcomes 93 linguistic and psycholinguistic features in total, but each and every feature are not closely related to detect troll comments. Therefore, we have used only the emotional psycholinguistic features as listed in Table II. Eight different parts of speech types are considered and

percentage of existence of each types are calculated using *Python Natural Language Toolkit (NLTK) Library* [15]. The percentages of slang words, colloquial words and sentiment words are counted using the ground-truth word databases computationally. SentiStrength estimates positive (1 to 5) and negative (-1 to -5) within limited range, where lower value means lower strengths and vice versa. The positive and negative strength values from the comments are extracted using the Java interface of SentiStrength. The feature vector contains 19 different numeric values which could be used for building a classification based predictive model. The predictive model could be considered as a binary classification problem as the comments could be determined as troll comment or not a troll comment.

### D. *Classification Algorithm based Predictive Model*

The extracted features are send to the classification algorithms to build a training model for prediction. The system would be 10-fold cross validated for testing. For our study, we have applied Multinomial Naïve Bayes (MNB), Decision tree (J48) and sequential minimal optimization (a fast algorithm for SVM) for binary classification problem. We have used Weka [19] for building the prediction models using these classifiers. Developed in University of Waikato, Waikato Environment for Knowledge Analysis (Weka) [19] is a collection of machine learning tools including many classifiers and attribute selection algorithms. The details about the classification algorithms could be found in [16], [17] and [18]. The performances of the prediction model could be illustrated from Table III.

TABLE III. PERFORMANCE MATRICS OF CLASSIFIERS

| Classifier | Precision | Recall | F-Measure | Accuracy (%) |
|---|---|---|---|---|
| MNB | **0.893** | **0.897** | **0.899** | **89.67%** |
| J48 | 0.875 | 0.840 | 0.849 | 84.18% |
| SMO | 0.800 | 0.800 | 0.799 | 80.00% |

From the performance table we can say, MNB has shown better performance in terms of f-score and accuracy, which is also evident in many natural language processing task [20].

### E. *Detecting Troll Comments*

Using the predictive model the system will detect which comment is a troll comment and which is not. These decision could be used for further detection of character assassination.

## V. FRAMEWORK FOR CHARACTER ASSASSINATION DETECTION

In this paper, we have presented the framework for detecting Character assassination based on the troll comment detection approach. The framework utilizes the decision coming from section IV. The framework first determines if the posted comment is a troll comment or not. If the comment is detected as a troll comment, then the personality traits of the victim user are determined using the psycholinguistic tools such as LIWC. We have used *myPersonality* dataset [13, 14], which contains around 10,000 Facebook status updates of 250 users. The dataset gives ground-truth personality scores and labels along with the status and social network features. We have trained using this popular dataset for further testing and recommendations.

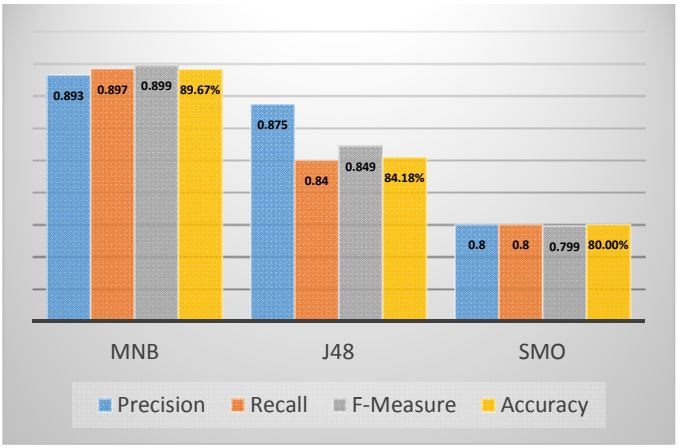
**Detecting Character Assassination**

**Steps:**

1. Input comments to the system.
2. Detecting troll comments using the proposed approach.
3. If the comment is detected as troll comment, determine the big five personality traits of the victim user. Otherwise, do nothing.
4. Based on the personality trait labels determine if the troll comment is for or against the actual personality of the victim.
5. If actual personality does not supports the troll comment, then detect the troll account and recommend block or ban to the victim user. Otherwise, do nothing.

## VI. EMPIRICAL ANALYSIS

In section III, the proposed approach shows better result for multinomial Naïve Bayes classifier (*89.67%*), than the traditional decision tree (*84.18%*) and SMO (*80.00%*). The comparative analysis between the proposed classifier algorithms could be depicted as in Fig. 2. The framework of detecting character assassination would have satisfactory results as it is utilizing the proposed method of detecting troll comments.




## VII. CONCLUSION

In this paper, we have presented a novel approach for detecting character assassination based on troll comment detection in the context of social media. The proposed troll comment detection approach is depending on the closely related features such as sentiment features, emotional affect, slang words, colloquial words etc. The novel approach of detecting gives satisfactory accuracy to be used for applications. The main contribution of this paper is to propose an approach to determine troll comments and utilizing the decision for detecting character assassination. The character assassination process involves the widely accepted BFFM. Context based recommender systems could be devised from the detection of character assassination such as blocking or banning the troll account for the victim user. The troll account holder will not be able to access to the victim users profile if he/she tends to post troll comments. This could be considered as one of the main future development scope to be developed on the basis of this papers proposed methods.

## REFERENCES


[1] M. M. Hasan, N. H. Shaon, A. A. Marouf, M. K. Hasan, H. Mahmud and M. M. Khan, "Friend Recommendation Framework for Social Networking Sites using User's Online Behavior", 18th IEEE International Conference on Computer and Information Technology (ICCIT), MIST, Bangladesh, 21-23 December, 2015.

[2] P. Howlader, K. K. Pal, A. Cuzzocrea and S. D. M. Kumar, "Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques", SAC '18 Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 339-345, Pau, france, April, 2018.

[3] T. Tandera, Hendro, D. Suhartono, R. Wongso and Y. L. Prasetio, "Personality Preddiction System from Facebook Users", 2nd International Conference on Computer Science and Computational Intellignece (ICCSCI), Bali, Indonesia, October, 2017.

[4] Y. R Tausczik and J. W Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods.", Journal of language and social psychology vol. 29, no. 1, pp. 24–54, 2010.

[5] G. Miller, "WordNet: A Lexical Database for English," Comm. ACM, vol. 38, no. 11, pp. 39–41, November, 1995.

[6] S. Dhuliawala, D. Kanojia and P. Bhattacharyya, "SlangNet: A WordNet like resource for English Slang", Language Resources and Evaluation Conference (LREC 2016).

[7] J. P. McCrae, I. Wood and A. Hicks, "The Colloquial WordNet: Extending Princeton WordNet with Neologisms" In: Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., Hellmann S. (eds.) Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol. 10318. Springer, 2017.

[8] V. Kaushal and M. Patwardhan, "Emerging trends in personality indentification using online social networks-A literature survey", ACM Transactions on Knowledge Discovery from Data, vol. 12, no. 2, Article. 15, January 2018.

[9] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", Proceedings of the Seventh conference on International Language Resources and Evaluation, Retrieved May 25, 2010.

[10] A. Esuli, and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining", Proceedings of Language Resources and Evaluation (LREC) 2006, Retrieved July 28, 2009.

[11] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text", Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544–2558, 2010.

[12] M. Thelwall, K. Buckley and G. Paltoglou, "Sentiment strength detection for the social Web", Journal of the American Society for Information Science and Technology, vol. 63, no. 1, pp. 163-173, 2012.

[13] M. Kosinski, S. Matz, S. Gosling, V. Popov and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.", American Psychologist. 70(6): pp. 543, February 2015.

[14] M. Kosinski, D. Stillwell and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior.", In Proceedings of the National Academy of Sciences of the United States of America (PNAS), pp. 5802-5805, 2013.

[15] S. Bird and E. Loper, "NLTK: The Natural Language Toolkit", Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, 2004.

[16] I. Rish, "An empirical study of the naive bayes classifier", In Proceedings of IJCAI-01 workshop on Empirical Methods in AI", pp. 41–46, Sicily, Italy, 2001.

[17] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology", IEEE Transactions on Systems, Man and Cybernetics, pp. 660–674, 1991.

[18] J. C. Platt, "Sequential minimal oprimization: A fast algorithm for training support vector machines", Technical REport MSR-TR_98_14, Microsoft Research, 1998.

[19] G. Holmes, A. Donkin and I.H. Witten, "Weka: A machine learning workbench", Proceddings of Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

[20] S. B. Kim, K. S. Han, H. C. Rim and S. H. Myaeng, "Some effective techniques for naive bayes text classification", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457–1466, November, 2000.