JOURNAL OF CRITICAL REVIEWS
ISSN- 2394-5125 VOL 7, ISSUE 16, 2020

CYBERBULLYING DETECTION IN ROMAN URDU LANGUAGE USING LEXICON BASED APPROACH

Kazim Raza Talpur [1], Siti Sophiayati Yuhaniz [2], Nilam Nur binti Amir Sjarif [3], Bandeh Ali [4]

[1]Razak Faculty of Technology & Informatics, UTM, Kuala Lumpur, 54100, Malaysia, talpurkazim@gmail.com
[2]Razak Faculty of Technology & Informatics, UTM, Kuala Lumpur, 54100, Malaysia, sohpia@utm.my
[3]Razak Faculty of Technology & Informatics, UTM, Kuala Lumpur, 54100, Malaysia, nilamnur@utm.my
[4]School of Computer Science & Statistics, Trinity College Dublin (TCD), Dublin, Dublin2, Ireland, bandehali@gmail.com
Corresponding author Email: talpurkazim@gmail.com

Received: 16 March 2020 Revised and Accepted: 19 June 2020

ABSTRACT: Nowadays, online social networks (OSNs) have become integral part of our daily life and online users of social media are massively growing. The increasing use of OSNs by users leads to large amount of user communication data. This study focuses on OSNs users who communicate in Roman Urdu (Urdu language written in English alphabets). Pakistan alone has over 44 million OSNs users who communicate in Roman Urdu. In this paper, we addressed the issue of cyberbullying behavior on Twitter platform, where users use Roman Urdu as medium of their communication. To the best of our knowledge, this is the first study addressing cyberbullying behavior in Roman Urdu. To address this issue, we developed supervised machine learning method and proposed a lexicon-based model with set of features derived from Twitter. An evaluation model shows that the developed model attained results with area under receiver operating characteristics curve (AUC) of 0.986 and f-measure of 0.984. These results indicate that the proposed lexicon-based method gives feasible solution for detecting cyberbullying behavior in Roman Urdu in OSNs. Finally, we compared results achieved with our proposed lexicon-based method and the results obtained from other well-known models. The comparison results show the significance of our proposed model.

KEYWORDS: Cyberbullying; online social networks (OSNs); supervised machine learning; Roman Urdu; Twitter

1. INTRODUCTION

With the emergence of internet and technology, online social networks (OSNs) have become significant aspect of our lives. It is an ongoing entertainment source. It enables us to maintain contact with each other by using a few taps and/or swipes in several apps. Even though people are at home or at work, they have become more social. In recent years, there have been an surge in the number of users with our smartphones and tablets on our social media platforms (Chaffey, 2019).

According to Worldwide digital report there are approximately 4,021 billion internet users, with around 3,196 billion users of social media and 5,135 billion users of mobile phones. Nonetheless, social media have their own challenges and problems. For instance, social media can include numerous anti-social behaviours, like cyberstalking and cyberbullying. These behaviours are now part of our lives and not only limited to young people, but everyone can suffer. Social media has been used as a new platform by cyber criminals to commit different types of Internet crime, in particular as phishing (Aggarwal et al., 2012), spamming (Yardi et al., 2010), malware of spread (Yang et al., 2012) and cyberbullying (Weir et al., 2011). With the recent development of online communication and social media (O'Keeffe & Clarke-Pearson, 2011), (Raisi & Huang, 2017), cyberbullying has emerged as a major issue that damages the lives of people along.

Cyberbullying can be defined as an individual or groups of users harassing other users using information and communication technology. Cyberbullying is also widely known as a grave national health issue (Xu et al., 2012) and identified as social threat (Özel et al., 2017), wherein victims show a substantially higher risk of suicide ideation (Sampasa-Kanyinga et al., 2014). In addition, cyberbullying is a considerably perpetual type of traditional forms of bullying with adverse consequences on victims. OSN websites (i.e. Twitter and Facebook) have become important tools of user life. Accordingly, these OSNs have become most frequent and common platform for cyberbullying harassment and victimization (Whittaker & Kowalski, 2015), and over the last few years their fame and rapid growth have been increased in terms of cyberbullying incidents (Bollen et al., 2011; Van Hee et al., 2018).

Twitter is an OSN service that allows users to exchange 280 tweet character messages (Agrawal & Singhal, 2019). Currently, Twitter includes around 500 million users and about 288 million active users to communicate each other and produce around 500 million tweets every day. However, this OSN website has become more important, actual connectivity communication platform (Kavanaugh et al., 2012). Research study showed that Twitter is turning into a "cyberbullying playground" and cyberbullying on Twitter network is rapidly increasing day by day due to large number of active users on this network (Chen et al., 2012a; Graham & Haarstad, 2014).

A. Cyberbullying in Pakistan

According to Geonews (news channel) in Pakistan more than 44 million users use social media networks such as Facebook, Twitter, Line, and Snapchat[1].

In addition, Pakistan Telecommunication Authority (PTA) facts shows that 162 million are cellular subscribers and 74 million are broadband subscribers[2]. Despite the fact, rapidly increasing social media users in Pakistan, give rise to many issues in society such as intellectual property theft, spams, phishing and other forms of social engineering. Notwithstanding that the above mentioned consequences of use of social media, cyberbullying is serious problem of Pakistan (Mohsin,2016). Moreover, recently published first online violence report by Geonews about cyberbullying, which showed that around 40% of women facing online harassment through internet[3].

B. Urdu Language

According to 2017 report of Ethnologue, Urdu, along with Hindi is the 3rd most commonly spoken language worldwide, with nearly 329.1 million native speakers, and 697.4 million total speakers[4]. Urdu is a national language in Pakistan and also widely spoken in India. There are around 11 million native Urdu speakers in Pakistan (A. Daud et al., 2017) and around 300 million native speakers live in India, UAE, United Kingdom and United States of America (Riaz, 2008). Urdu is written right-to left in an extension of the Persian alphabet, which is itself an extension of the Arabic alphabet. The family tree of Urdu language can be presented as: Indo-European→Indo–Iranian→Indo-Aryan→Urdu. Urdu is originated from Persian and Arabic and has similarities to most languages of South Asia. For instance, similarity in respect of: lack of capitalization, lack of small and capital words as well as free word order characteristic (A. Daud et al., 2017).

There are 39 basic letters and 13 additional characters in the Urdu language. It is written from right to left and is closely related to the Arabic and Persian alphabets, but also contains some sounds from Sanskrit[5]

C. Roman Urdu

Large number of Urdu speakers are using Roman script (for example English language alphabets) called Roman Urdu, for writing in Perso-Arabic script and in social media (Rafae et al., 2015). The people of Pakistan prefer Urdu writing in Roman Urdu and the effects of Roman Urdu are to decrease the capability of writing English and Urdu (Masroor et al., 2019). There is no standard for spelling the word in Roman Urdu. A single word can be written in different forms with diverse spelling by different people as well as by same person. Specifically, there is no particular mapping between Urdu letters for vowel sounds and the related roman letters (Bilal et al., 2016).

Roman Urdu is deficit of standard lexicon and generally a given word can be written with many spellings , e.g., the word zindagi [life] can also be written as zindagy, zindagee, and zindagi. Explicitly, the following standardization issues arise: (1) words with diverse spellings (the above example), (2) words with similar spellings but are gramatically different (e.g., "bahar" can be used for spring and outside and (3) spellings that match words in English (e.g, "had" means limit in Urdu for the English word "had"). These discrepancies cause a problem of data sparsity in basic natural language processing (NLP) tasks such as Urdu word segmentation, part of speech tagging, spell checking, machine translation, etc. (Rafae et al., 2015). However, in digital world Roman Urdu language is very famous while users are using OSNs such as Twitter, Facebook, Instagram, and mobile message. In Roman Urdu language writing style users are using English language alphabets and there is no standardized spellings in place (A. Daud et al., 2017).

The remainder of this paper is organized as follows. Section 2 presents the related work. Research methodology is given in section 3. In addition, section 3 also presents our novel lexicon-based feature engineering technique. Section 4 includes experiments and results. Discussion of observed results is given in section 5. Finally, section 6 concludes this work.

[1] www.geo.tv/latest/131187-Over-44-million-social-media-accounts-in-pakistan
[2] www.pta.gov.pak/en/telecom-indicators
[3] www.geo.tv/latest/143464-40-of-women-face-harassment-on-internet-says-pakistan-first-online-violence-study
[4] https://www.ethnologue.com/guides/ethnologue200
[5] http://www.bbc.co.uk/languages/other/urdu/guide/alphabet.shtml

D. Previous work

Daud et al.,(M. Daud et al., 2015) proposed an application RUOMiS (Roman Urdu Opinion Miner System), an automatic opinion mining system that mine and translate the Roman-Urdu and/or Romanagari reviews and provide the rating of the products based on users comments. The research helps the non-Urdu speaking users in selection of product by translating the comments, finding their polarity and then giving the rating of the product.

Sentiment mining in Roman Urdu Language has been done by (Bilal et al., 2016).Classification techniques such as Naive Bayes, Decision Tree and KNN were used for text classification. Training labeled dataset which are collected from blog contained 300 opinions including 150 positive and 150 negative opinions Navies Bayes get higher results such as precision, recall, and F-measure as compared to KNN and Decision Tree.

Bi-lingual classification method using NLP and sentiment analysis was proposed for English and Roman-Urdu tweets (Urdu language messages written using English alphabets), by (Javed & Afzal, 2014), (Javed et al., 2014). Tweets were collected from five major cities in Pakistan (Islamabad, Lahore, Karachi, Peshawar and Quetta) and belonged to various political parties. The collection of tweets expressed public opinions and views about different political parties. The performance of language classifier was measured using specificity, recall, precision, accuracy, error rate and F-measure.

(Afzal & Mehmood, 2016) performed spam classification for Roman Urdu tweets, collected from different cities of Pakistan. Various classifiers namely Naive Bayes Multinomial, Liblinear, LibSVM, DMNBText and J48 were used and to measure performance of these classifiers, accuracy and AUC were used.

In addition to these, (Mehmood et al., 2015) analysed automatic spam filtering, in Roman Urdu mobile text messages. Naïve Bayes Multinomial, DMNBText, LibSVM, Liblinear and Sequential Minimal Optimization (SMO) were used as algorithms in machine learning methods. However, performance was measured using accuracy and AUC.

The above-mentioned studies in Roman Urdu Language are limited to areas such as business development, marketing, product development, product feedback, public opinions and views and translation. A wide search on available articles was performed and no preceding work for cyberbullying detection in Roman Urdu language text and comments was found.

E. Lexicon based techniques

These techniques are based on the simple Bag-of-Words (BoW) approach. In this approach, a corpus of delicate, abusive, and unpleasant words is created (Pawar & Raje, 2019).

2. RELATED WORK

Van Hee et al., worked on automatic cyberbullying detection in social media text by exhibiting the posts written by bullies, victims, and bystanders of online bullying. They used linear support vector machines (SVM) and performed a series of experiments on binary classification to determine automatic cyberbullying detection. English and Dutch corpus were created after collecting data from ASKfm. For the automatic detection of cyberbullying, binary classification experiments using SVM implemented in LIBLINEAR was performed by operating Scikit-learn, a machine learning library for Python. The results report AUC scores more robust to data imbalance than recall, precision, and F score (Van Hee et al., 2018).

Duwairi detected sentiments from dialectical Arabic texts after applying two detection methods. First, dialectical words were translated into Modern Standard Arabic (MSA), then detected according to MSA lexicon and second method was detecting dialectical lexicon. Classifiers like Naïve Bayes (NB) and SVM were used to detect negative and positive polarities. The dialect lexicon presented a positive impact on the Macro-Precision, Macro-Recall and F-Measure. Furthermore, results showed that the F-measure of the Positive and Negative classes significantly improved by dialect lexicon in contrast to the Neutral class (Duwairi, 2015).

Cybernetic harassment in OSNs, in Spanish language was proposed by (Mercado et al., 2018). In this study, sentiment analysis techniques, such as bag of words, removal of signs and numbers, tokenization and stemming were performed. The database of Agustn Gravano (SDAL), which was a lexicon of 2880 words, from the Faculty of Exact and Natural Sciences of the University in Buenos Aires, Argentina, was used.

(Del Bosque & Garza, 2014) proposed automatically mapping of a document with an aggressiveness score for cyberbullying and explored different approaches including lexicon-based, supervised, fuzzy, and statistical approaches. To do so three different lexicons were used in the study which include swear words (named NS) was extracted from the noswearing.com site, ANEW lexicon, which stands for “Affective Norms for English Words” and SentiWordNet. The score of these three lexicons and document length (number of words), number of offensive words and frequency of the word "you" were employed as set of features. The overall results with the best approach (lowest Mean Squared Error, MSE) was the 2-attribute linear regression, followed by the 3-attribute neural network, the fuzzy system, the NS lexicon, SentiWordNet, and lastly ANEW.

Chen et al., work's was based on identifying offensive contents in social media applying Lexical Syntactical Feature (LSF) approach. The dataset was retrieved from YouTube text comments from posted in reaction to 18 videos which include Music, Entertainments, Films, Autos, Educations, Comedies, News, Gaming, Animals, Style, Non-profits, Sciences, and Sports. Each text comment includes a user id, a timestamp and text content. The dataset includes comments from 2,175,474 different users. Machine learning techniques such as NB and SVM were used to perform the classification and to evaluate the performance of LSF, standard evaluation metrics i.e., precision, recall, and f-score were used. The authors claimed to improve the traditional machine learning methods by using lexical features to detect offensive languages as well as incorporating structure features, style features and context-specific features to better predict user's potentiality to disseminate offensive content in social media (Chen et al., 2012b).

The authors Dinakar et al. (2011) focused on detection of textual cyberbullying. They develop a corpus of comments from YouTube videos involving sensitive topics related to race & culture, sexuality and intelligence.

Four supervised learning algorithms namely; Nave Bayes, Rule-based, Jrip, Tree-based J48, and SVM were applied classify these topics. To build a corpus, they used different text mining techniques such as TF-IDF, POS, n-grams tokenizer, and also the list of profane words, the Ortony lexicon for negative affects (Dinakar et al., 2011).

Similar lexicon-based work is also done in foreign languages, some studies are highlighted here. (Gómez-Adorno et al., 2018) presented a method to detect aggressive tweets in Spanish. In their method, logistic regression classifier was trained on linguistic patterns, aggressive words lexicon, and several types of n-grams with oversampling technique to balance the classification distribution. Classifier achieved F-measure score of 42.85 on aggressive class on training data but method poorly performed on testing data.

Another lexical approach was applied on Indonesian language to detect cyberbullying in Twitter.

In this study, data mining technique was used to mine data in the database and then data was analysed with association rule and FP-Growth methods to find trends and patterns in collected dataset from Twitter. Both methods showed similar results to find frequent items in the database (Margono et al., 2014a).

Arabic language social media comments from YouTube and Twitter were also analysed using lexicon approach, based on corpus of cyberbullying and aggressive words. Comments were classified with weighted function into three categories such that; mild, medium, and strong. Method showed significance in results by identifying most of cyberbullying comments (Mouheb et al., 2019).

Table1: Summary of cyberbullying detection in other languages

Studies	Area	Languages
1 (Emon et al., 2019)	Social Media Sites	Bengali
2 (Özel et al., 2017)	Twitter and IG	Turkish
3 (Pawar & Raje, 2019)	Twitter and Reviews	Hindi & Marathi
4 (Hussain et al., 2018)	Social Websites	Bangla Text (Bengali)
5 (Gómez-Adorno et al., 2018)	Twitter	Mexican and Spanish
6 (Margono et al., 2014b)	Twitter	Indonesian
7 (Febriana & Budiarto, 2019)	Twitter	Indonesian
8 (Mouheb et al., 2018)	Twitter & YouTube	Arabic
9 (Mouheb et al., 2019)	Twitter	Arabic
10 (Van Hee et al., 2015)	Social Network site	Dutch
11 (Bai et al., 2018b)	Twitter	German

12	**(Bai et al., 2018a)**	Social Media Sites	Italian

3. RESEARCH METHODOLOGY

This section comprehensively discusses the methods utilized in cyberbullying detection in Roman Urdu from OSN. All steps in this research are presented in figure 1 and explained in the following sections. In this study, supervised machine learning method is used to train classifier to detect cyberbullying behaviour in tweets written in Roman Urdu.

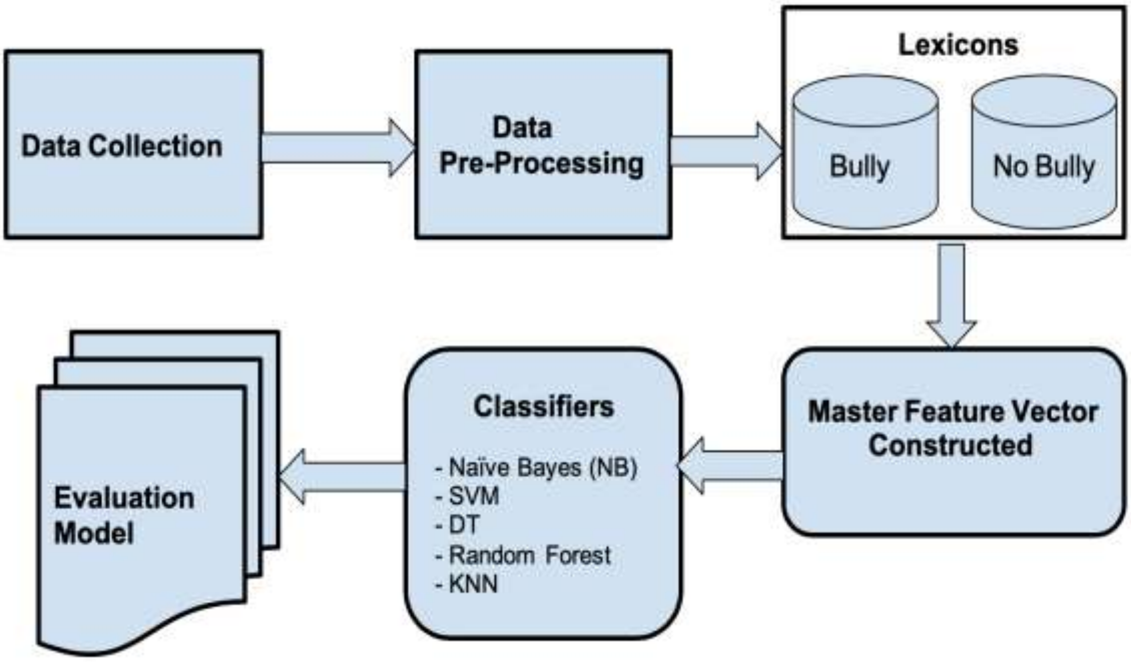

Figure 1. Proposed framework

A. Document pre-processing;

One of the critical tasks in the text mining is converting text from unstructured format into structured form, as most of the time text co-occurs with a lot of unnecessary data such as tags, anchor text, and other irrelevant features. Therefore, it is critically important to pre-process input data before any other operation. Generally, text pre-processing steps can be as follows;

1. Tokenization: is the process of breaking the text corpus into words (most commonly), phrases, or other meaningful elements, which are then called tokens.
2. Stop words removal: These are the words that do not add any value to the context or meaning in the text. By removing these words, we can focus on important words or phrases to improve accuracy. The idea is simply to remove words that occur commonly across all the documents in the corpus.
3. Stemming and Lemmatization: Stemming refers to reduction of a word into its root form. For example, rain, raining, rained, rainfall all reduces to common root word "rain". Lemmatization on the other hand is a more advanced form of stemming that attempts to group words based on their core concept or lemma. Lemmas are the base forms of words that are used to key the word in a dictionary. For example, trouble, troubling, troubled, troubles all group into lemma “trouble”.
4. In this study, we used Nature Language Toolkit (NLTK[6]) python library for text processing to perform Tokenization, stop words removal, stemming, and lemmatization;

[6] http://www.nltk.org/

In this study, we used Nature Language Toolkit (NLTK[7]) python library for text processing to perform Tokenization, stop words removal, stemming, and lemmatization;

B. Data Collection and Labelling;

Twitter users from all over the world collectively produce large number of tweets that posted on Twitter network. The Twitter network, application program interface (API) allows the researcher to extract publicly tweets. Each tweets contains large-scale information(Kwak et al., 2010) such as username, user ID, biography of user, screen name of user, user URL, user account creation information of data, tweet of text i.e. the main tweet of text information have about thought, emotions, behaviours, and other user silent information (Eichstaedt et al., 2015), creation time of tweet, unique ID of tweet's, tweet of language, user tweets number, user favourites of number, user number of following, user mentions of number, user amount of re-tweets, user location (geo-location) and the application of user that sent the tweet (Bollen et al., 2011; Eichstaedt et al., 2015; Preoţiuc-Pietro et al., 2015).

In first step of this research study, language specific tweets were collected from Twitter using developer API. Our dataset contained 2 million tweets in Roman Urdu. It was planned to extract real time tweets involving two types of cyberbullying and non-cyberbullying. We extracted only publicly available content via Twitter API and according to Twitter network privacy and polices to avoid any privacy breach.

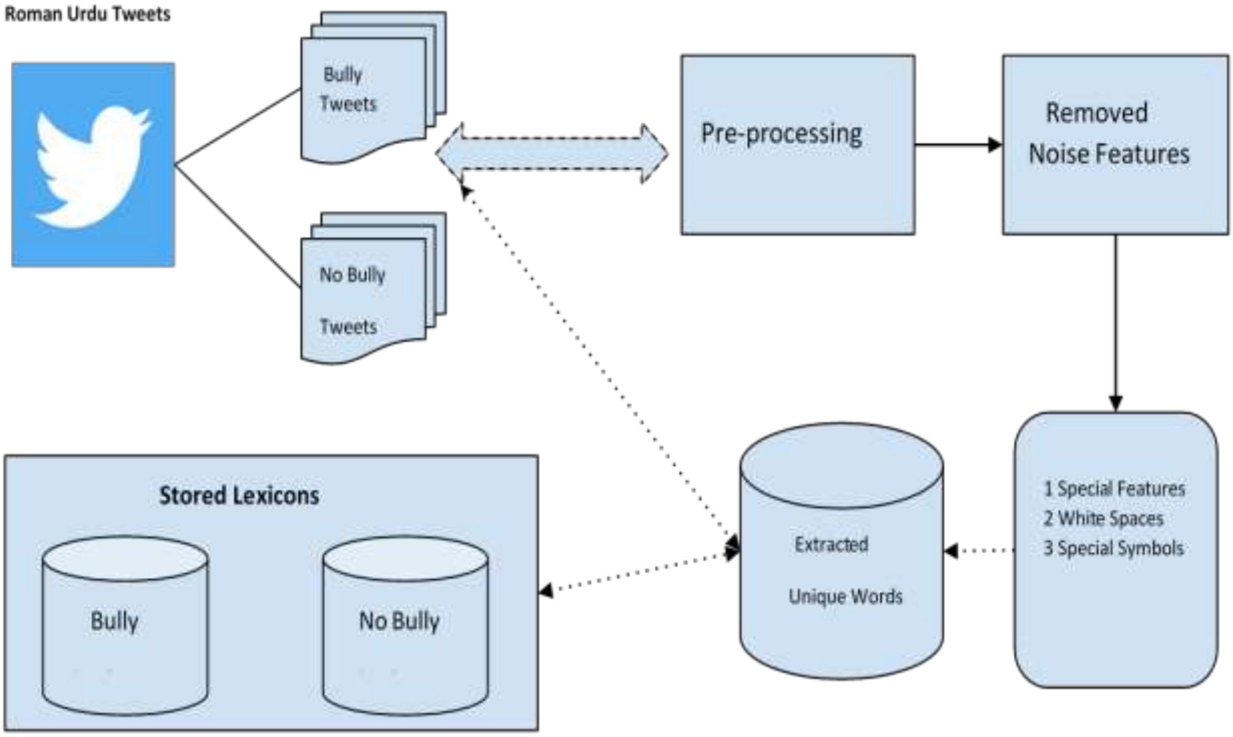

Figure 2. Process of proposed lexicon-based approach

We then randomly selected 18,000 tweets from our collected dataset for labelling. In this research, Roman Urdu tweets were labelled with the help of three native Urdu language experts. Furthermore, these experts were oriented about the slang words, abbreviations and acronyms which is commonly used in OSNs. Tweets were considered cyberbullied when at least two of the labelers agreed to classify same tweet for cyberbullying behaviour. Tweets were removed from the dataset if labelers did not agree on tweet classification. Our final dataset contained 17,357 tweets, from which, 16,978 were classified as non-cyberbullied and 379 as cyberbullied tweets.

C. Feature Engineering and Master Feature Creation

The machine learning algorithm cannot learn the classification rules from raw text. These algorithms need numerical vectors to learn classification rules. Therefore, the raw text needs to be converted into numerical vectors through feature engineering approaches. In literature, there are several feature engineering approaches available to convert raw text into numerical vectors. These include, Bag of Words, n-gram and Word2Vec.

[7] http://www.nltk.org/

However, our experimental results showed that these techniques yielded the AUC not greater than 80%. Thus, to obtain better results compared to existing baseline feature engineering techniques and to convert our collected tweets into numerical vectors, we employed a novel lexicon-based feature engineering approach

Figure. 2. shows the functionality of our proposed lexicon-based approach. As shown here, we first separate the bullying and non-bullying tweets. Afterwards, from each category, we applied some pre-processing techniques to remove noisy features. In pre-processing techniques, we initially removed white spaces and special symbols using regular expression notations. Furthermore, we extracted the unique words from bullying and non-bullying categories and stored the extracted unique words into two different lexicons namely, bullying lexicons and non-bullying lexicons. Afterwards, the stop-words were removed from each lexicon because of their frequent availability in both lexicons.

After the preparation of bullying and non-bullying lexicons, we created the master feature vector where each tweet was converted into numerical vector. In this master numerical vector, each row represents one instance of tweet and have two columns representing each lexicon. For converting each tweet into numerical vector, we took each tweet, removed the white spaces from the tweet, tokenize the tweet into words, removed the stop words, and finally compute the words matched from bullying and non-bullying lexicons using below equation.

S = number of words matched from lexicon / total words in lexicon

$$S = \frac{f_m}{f_t} \quad (1)$$

The main function of master features creation was if a word occurs in positive word list, it gets +1 polarity value. If it is found in the negative word list, -1 is assigned to it. If the same word is repeated twice, it is assigned polarity only once i.e. either +1 or -1 depending on the tweet word, whether it is positive word or negative word.

The main steps of the master feature vector are as follows;

1. Words in each tweet are searched in positive and negative list.
2. Words are assigned polarities according to specific rules.
3. When the end of the tweet is reached, the polarities assigned to the words in each tweet are added.
4. Total polarity score is generated which shows the total number of positive and negative polarity score for each tweet.
5. If positive score is greater than negative score, tweet is classified as '1' for cyberbullied, and '0' for non-cyberbullying tweet.
6. Similarly, if negative score is greater than positive score, tweet is classified as '0' for non-cyberbullying and '1' for cyberbullied

After creation of master feature vector, we fed this master feature vector as an input to five different machine learning algorithms (namely, SVM, NB, RF, DT, KNN) to construct the classification model. The detail can be found in subsequent section.

D. Machine Learning Algorithms

Most significant step of text classification process is choosing the best classifier. The features extracted from the tweets were applied to develop a model for detecting cyberbullying in Roman Urdu. We tested various types of machine learning methods and chose the best classifiers namely; Naïve Bayes (NB), Support Vector Machine (SVM), KNN, Decision Tree and Random Forest. According to Sandhya, mostly used classification algorithms are Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest and Nearest Neighbour (Sandhya, 2019).

(a) **Naïve Bayes (NB)**

NB algorithms is collection of algorithms and works with independent assumptions which is based on Bayes theorem. It is one of the powerful and easy-to-train classifiers that determine the probability of an outcome given a set of conditions using Bayes' theorem. Moreover, NB is one of the well-known supervised machine leaning algorithm, and main function of this algorithm is to maximize the posterior probability which is given in training data phase to develop a decision model for new data (Nair et al., 2019).

NB is considered is one of the most efficient and effective inductive learning in the area of machine learning (Zhang, 2004). This algorithm is easy to train with no complicated iterative parameter estimation which is makes it specifically convenient for large dataset (Saravanaraj et al., 2016) and it has been widely used classifier in various research studies on OSNs (Kovoor et al., 2018; Özel et al., 2017; Saravanaraj et al., 2016) . In this study, we used basic version of NB for document features and classification.

***(b)* Support Vector Machine (SVM)**

SVM is one of the powerful and versatile supervised machine learning algorithms which can perform linear and nonlinear classification and regression problems. It is based on statistical leaning theory (Vapnik, 2013) and main function of SVM is to find out hyperplane in the dimension-space which is clearly classifies into data points (Nair et al., 2019). SVM is well suited for outlier detection and has been widely used in cyberbullying detection and online harassment (Kovoor et al., 2018; Özel et al., 2017; Sheeba & Devaneyan, 2017; Yin et al., 2009).

***(c)* K-Nearest Neighbours (KNN)**

KNN algorithm is a non-parametric method which is widely applied in various application and includes text recognition, medical diseases diagnosis and web mining (Rajeshkanna et al., 2019). KNN can be described as lazy learning algorithm as it applies k nearest items surrounding a particular data point and tries to classify this data points into one of the output labels based on its k nearest data points. The k in KNN is the number of nearest neighbours to be considered. It is simplest machine learning algorithm as it does not make any assumption for underlying distribution of data (Nair et al., 2019) and it mainly compares the objects to be predicted with their KNNs to establish their own categories (Lee et al., 2018).

***(d)* Decision Trees (DT)**

DT is one of the most famous machine learning algorithm and commonly useful for solving classification problem in field of data mining (Farid et al., 2014). DT can handle both numerical data for regression problems and categorical data for classification problems.

The main advantage of this algorithm includes; it can handle training data with missing values and are simple to understand and interpret. It shows graphical representation of possible options for decision and events (Vallabhaneni, 2019). Generally, the main function of DT is to find out optimal decision tree through reducing the error of generalization.

***(e)* Random Forest (RF)**

RF is a supervised ensemble machine learning algorithm which is mostly used for classification and regression problems (Babbar et al., 2019; Nair et al., 2019; Saravanaraj et al., 2016). RF constructs several decision trees and combine them together to achieve more correct stable predication.

It is also convenient and easy to use machine learning algorithm that produces best and accurate results. Similar to DT algorithms, RF can handle training dataset with missing value but at large scale and still provides better accuracy.

E. Performance Evaluation

In supervised machine learning, there are several evaluation measures. Different evaluation measure different characteristics of model performance. Majority of these evaluation measures are built on confusion matrix (Table II), which identifies the classification of correctly and incorrectly instances. Generally, one of the most used machine learning measure is Accuracy, which does not distinguish the correctly classified labels for different classes:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

Table2: Confusion Matrix

Class Recognized as	**Positive**	**Negative**
Positive	True Positives - TP	False Negatives – FN
Negative	False Positives - FP	True Negatives - TN

Measures that evaluate the classifiers performance for different classes are Sensitivity and Specificity. These measures are often used in medical research which involves visual data.

In information retrieval, text classification and natural language processing focuses usually on importance of one class, which can be either positive or negative class. The number of instances belonging to one class in these areas of applications is substantially lower than the total instances (Sokolova et al., 2006). In such instances, measure of choice on single class are as follows;

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Precision is a function of true positives and instances incorrectly classified as positives (false positives). Recall is a function of correctly classified instances (true positives) and its incorrectly examples (false negatives). The F-measure on other hand is the combination of precision and recall, which is the weighted average of precision and recall.

Another widely used measure ROC (AUCROC or AUC) curve is a graphical representation of the trade-off between false-positive and false-negative rates for selected instances obtained from a test data. AUC is usually used when dataset is imbalanced, the higher the AUC the better the performance (He & Ma, 2013).

In this study, we used AUC as our main metric due to the nature of imbalanced data and we report weighted F-measure as reference measure. Weighted F-measure here is not harmonic mean of precision and recall but rather the sum of all measures whereby each weight is given according to the number of instances with that particular class label.

4. RESULTS AND DISCUSSION

This section presents the experimental results obtained from three well-known text classification techniques namely, bag-of-words (BOW), n-gram (trigram), Word2Vec and our proposed lexicon-based approach as to improved version approach in comparison to baseline approaches for detecting cyberbullying behaviour in Roman Urdu tweets.

All the experiments of three baseline approaches results given in the table 4 and proposed Roman Urdu language lexicons-based approach results are given in the table 3.

Table 3: Proposed Lexicon-based results

Approach	Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
Lexicon	NB	0.971	0.865	0.964	0.971	0.967	0.686	97.13%
	SVM	0.985	0.49	0.983	0.985	0.983	0.747	98.45%
	IBK	0.984	0.532	0.982	0.984	0.982	0.984	98.41%
	J48	0.984	0.54	0.982	0.984	0.982	0.859	98.38%
	RF	0.985	0.47	0.983	0.985	**0.984**	**0.986**	98.47%

Table 3: Results for baseline approaches

Approach	Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy
BOW	NB	0.934	0.698	0.964	0.934	0.948	0.708	93.40%
	SVM	0.977	0.955	0.962	0.977	0.968	0.511	97.68%
	IBK	0.961	0.873	0.961	0.961	0.961	0.542	96.14%
	J48	0.977	0.906	0.966	0.977	0.969	0.575	97.71%
	RF	0.978	0.97	0.968	0.978	**0.968**	**0.775**	97.82%
n-gram	NB	0.919	0.719	0.963	0.919	0.939	0.688	91.92%
	SVM	0.977	0.963	0.962	0.977	0.968	0.507	97.74%
	IBK	0.96	0.888	0.96	0.96	0.96	0.535	95.96%
	J48	0.977	0.947	0.964	0.977	0.968	0.542	97.74%
	RF	0.978	0.968	0.966	0.978	**0.968**	**0.713**	97.80%
Word2Vec	NB	0.958	0.852	0.961	0.958	0.96	0.68	95.79%

SVM	0.978	0.963	0.968	0.978	0.968	0.748	97.82%
IBK	0.955	0.899	0.959	0.955	0.957	0.531	95.52%
J48	0.973	0.929	0.961	0.973	0.966	0.511	97.29%
RF	0.977	0.942	0.964	0.977	**0.968**	**0.698**	97.71%

As shown in the Table 3, the proposed Roman Urdu language lexicon-based model performed better than BOW, N-gram and Word2vec approach. Random Forest achieved the best performance among all classifiers in baseline settings (bag-of-words, n-gram and Word2Vec) and in proposed lexicon-based approach. However, among all three baseline approaches BOW showed the best result compare to n-gram approach and Word2Vec. AUC in all baseline approaches for all selected classifier varied between 0.507 to 0.775 in BOW and in n-gram and Word2vec approach with f-measure varied between 0.939 to 0.969.

Whereas, AUC in proposed approach significantly improved and varied between 0.686 to 0.986, whereby Naïve Bayes showed the lowest AUC and best performing AUC among all classifier in proposed approach was Random Forest. Random Forest AUC score jumped from (baseline 1) 0.775 to 0.986 in proposed model and f-measure slightly improved from 0.968 to 0.984 (Figure 3).
It is worth noticing that false positive rate in all baseline approaches for all selected classifier were very high 0.698 to 0.97 compare to proposed approach where for most of classifiers false positive rate varied between 0.47 to 0.54 except Naïve Bayes classifier which had higher false positive rate of 0.865 (Figure 4).

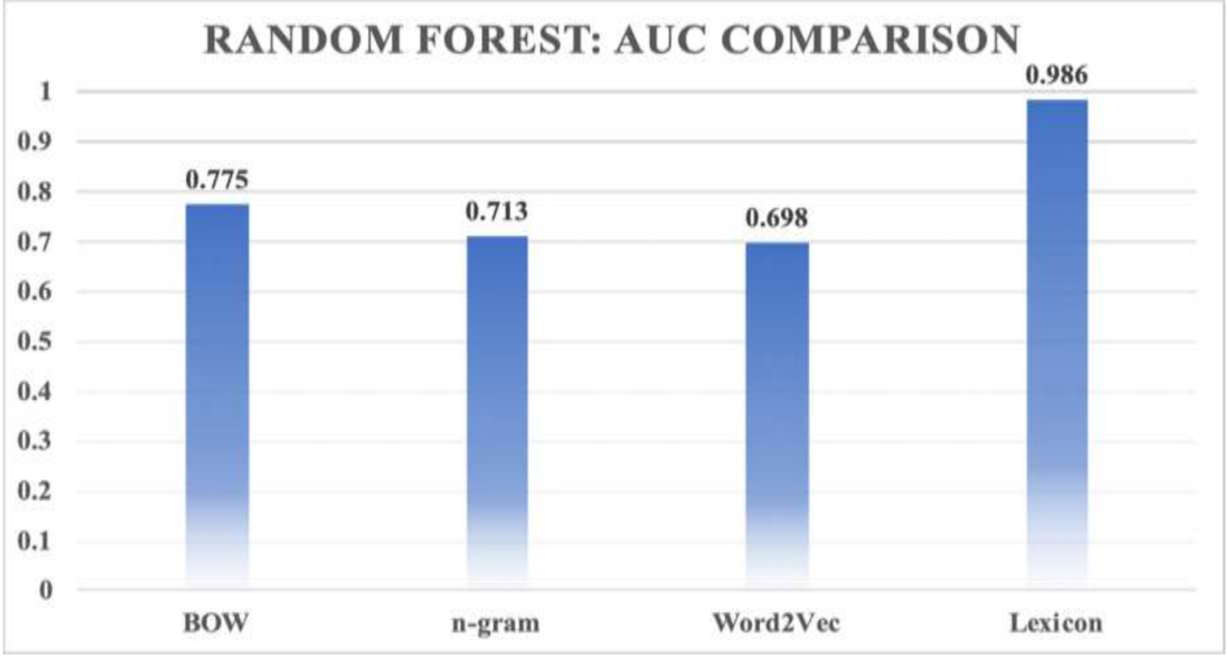

Figure 3. Comparison of AUC with baseline and proposed approach

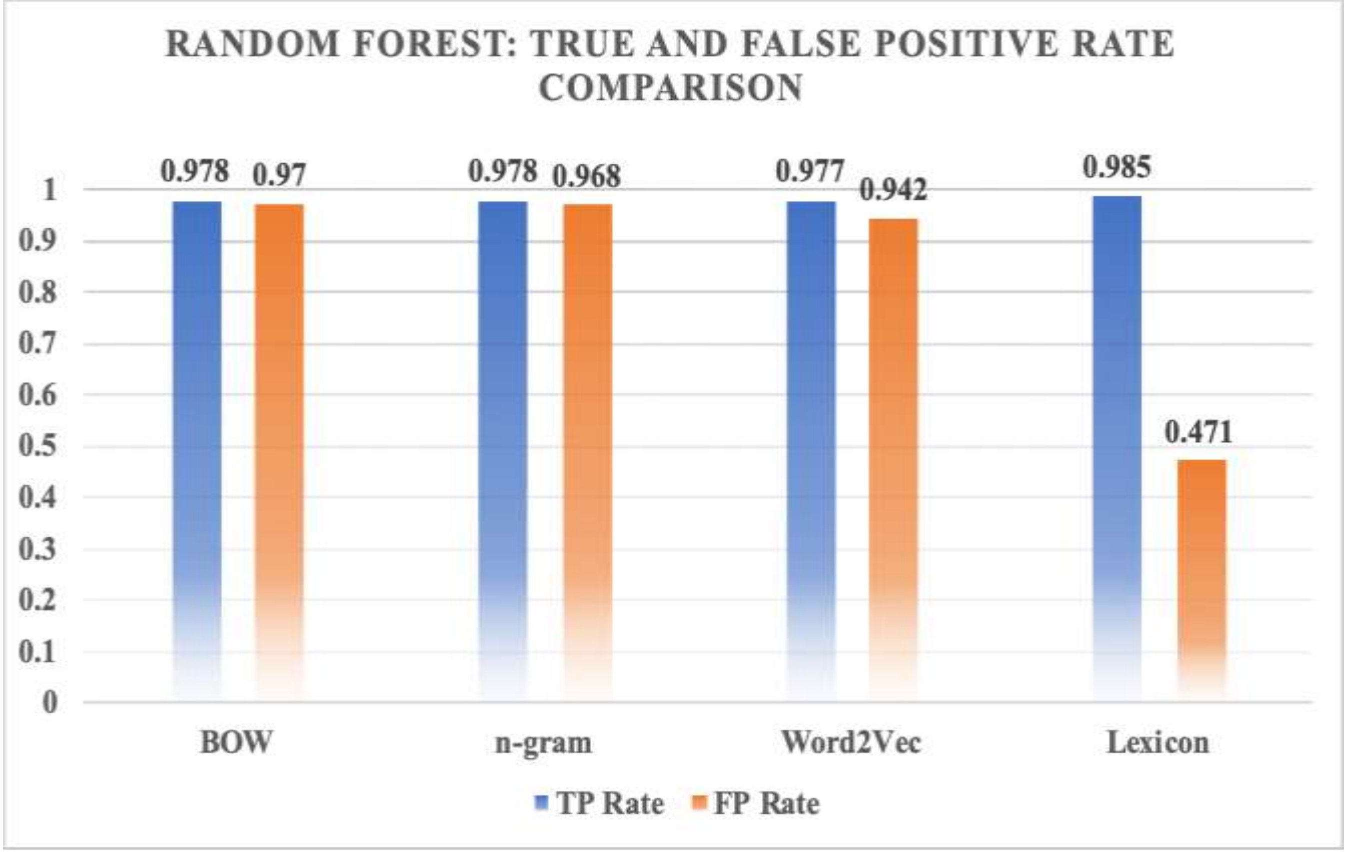

Figure 4. Comparison of True Positive rate with baseline and proposed approach

True positive rate seems to have reasonable rate in both baseline approaches. Furthermore, true positive rate slightly improved in proposed approach. The experimental results indicate that features such as, user followers, friends, user favourites, tweet time, tweet favourited, tweet retweeted, number of time statuses that has been updated since joining, and user duration for joining twitter barely affected baseline classifier performance. However, classifier performance significantly improved when aforementioned features are applied with our lexicon-based approach.

5. FUTURE STUDY

We believe that our Lexicon-based model can applied into any online social network datasets to detect cyberbullying instances in as efficient way. in addition, our model can also be deployed within any online social media networks to assist social media services for cyberbullying detection instances. Furthermore, there are several future directions in the area of cyberbullying detection, most of research studies only focused on English language, but there are various international and regional languages and mostly online users communicate each other in online social networks. However, it is very important to develop a models and test on large datasets to detect cyberbullying messages. Moreover, another possible future direction to increase features derived from online social networks and applied into cyberbullying detection model.

.

6. REFERENCES

[1] Afzal, H., & Mehmood, K. (2016). Spam filtering of bi-lingual tweets using machine learning. Advanced Communication Technology (ICACT), 2016 18th International Conference On, 710–714.

[2] Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: Automatic realtime phishing detection on twitter. 2012 ECrime Researchers Summit, 1–12.

[3] Agrawal, T., & Singhal, A. (2019). An Efficient Knowledge-Based Text Pre-processing Approach for Twitter and Google+. International Conference on Advances in Computing and Data Sciences, 379–389.

[4] Babbar, S., Kulshrestha, S., Shangle, K., Dewan, N., & Kesarwani, S. (2019). Improvization of Arrhythmia Detection Using Machine Learning and Preprocessing Techniques. In Applications of Artificial Intelligence Techniques in Engineering (pp. 537–550). Springer.

[5] Bai, X., Merenda, F., Zaghi, C., Caselli, T., & Nissim, M. (2018a). RuG@ EVALITA 2018: Hate Speech Detection In Italian Social Media. EVALITA 2018.

[6] Bai, X., Merenda, F., Zaghi, C., Caselli, T., & Nissim, M. (2018b). Rug at germeval: Detecting offensive speech in german social media. 14th Conference on Natural Language Processing KONVENS 2018, 63.

[7] Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences, 28(3), 330–344.

[8] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

[9] Chaffey, D. (2019). Global social media research summary 2019. Smart Insights. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[10] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012a). Detecting offensive language in social media to protect adolescent online safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 71–80.

[11] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012b). Detecting offensive language in social media to protect adolescent online safety. Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 71–80.

[12] Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: A survey. Artificial Intelligence Review, 47(3), 279–311. https://doi.org/10.1007/s10462-016-9482-x

[13] Daud, M., Khan, R., & Daud, A. (2015). Roman Urdu opinion mining system (RUOMiS). ArXiv Preprint ArXiv:1501.01386.

[14] Del Bosque, L. P., & Garza, S. E. (2014). Aggressive Text Detection for Cyberbullying. In A. Gelbukh, F. C. Espinoza, & S. N. Galicia-Haro (Eds.), Human-Inspired Computing and Its Applications (pp. 221–232). Springer International Publishing. https://doi.org/10.1007/978-3-319-13647-9_21

[15] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. Fifth International AAAI Conference on Weblogs and Social Media.

[16] Diplomat, M. M., The. (n.d.). The Cyber Harassment of Women in Pakistan. The Diplomat. Retrieved October 19, 2018, from https://thediplomat.com/2016/04/the-cyber-harassment-of-pakistans-women/

[17] Duwairi, R. M. (2015). Sentiment analysis for dialectical Arabic. 2015 6th International Conference on Information and Communication Systems (ICICS), 166–170. https://doi.org/10.1109/IACS.2015.7103221

[18] Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., & Sap, M. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological Science, 26(2), 159–169.

[19] Emon, E. A., Rahman, S., Banarjee, J., Das, A. K., & Mittra, T. (2019). A Deep Learning Approach to Detect Abusive Bengali Text. 2019 7th International Conference on Smart Computing & Communications (ICSCC), 1–5.

[20] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications, 41(4), 1937–1946.

[21] Febriana, T., & Budiarto, A. (2019). Twitter Dataset for Hate Speech and Cyberbullying Detection in Indonesian Language. 2019 International Conference on Information Management and Technology (ICIMTech), 1, 379–382.

[22] Gómez-Adorno, H., Enguix, G. B., Sierra, G., Sánchez, O., & Quezada, D. (2018). A Machine Learning Approach for Detecting Aggressive Tweets in Spanish. IberEval@ SEPLN, 102–107.

[23] Graham, M., & Haarstad, H. avard. (2014). Transparency and development: Ethical consumption through Web 2.0 and the internet of things. Open Development: Networked Innovations in International Development, 79.

[24] He, H., & Ma, Y. (2013). Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons.

[25] Hussain, M. G., Al Mahmud, T., & Akthar, W. (2018). An Approach to Detect Abusive Bangla Text. 2018 International Conference on Innovation in Engineering and Technology (ICIET), 1–5.

[26] Javed, I., & Afzal, H. (2014). Creation of bi-lingual Social Network Dataset using classifiers. International Workshop on Machine Learning and Data Mining in Pattern Recognition, 523–533.

[27] Javed, I., Afzal, H., Majeed, A., & Khan, B. (2014). Towards creation of linguistic resources for bilingual sentiment analysis of twitter data. International Conference on Applications of Natural Language to Data Bases/Information Systems, 232–236.

[28] Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., Natsev, A., & Xie, L. (2012). Social media use by government: From the routine to the critical. Government Information Quarterly, 29(4), 480–491.

[29] Kovoor, B. C., Nandakumar, V., & Sreeja, M. U. (2018). CYBERBULLYING REVELATION IN TWITTER DATA USING NAÏVE BAYES CLASSIFIER ALGORITHM. International Journal of Advanced Research in Computer Science, 9(1).

[30] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web, 591–600.

[31] Lee, P.-J., Hu, Y.-H., Chen, K., Tarn, J. M., & Cheng, L.-E. (2018). Cyberbullying Detection on Social Network Services. PACIS, 61.

[32] Margono, H., Yi, X., & Raikundalia, G. K. (2014a). Mining Indonesian cyber bullying patterns in social networks. Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147, 115–124.

[33] Margono, H., Yi, X., & Raikundalia, G. K. (2014b). Mining Indonesian cyber bullying patterns in social networks. Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147, 115–124.

[34] Masroor, H., Saeed, M., Feroz, M., Ahsan, K., & Islam, K. (2019). Transtech: Development of a novel translator for Roman Urdu to English. Heliyon, 5(5), e01780. https://doi.org/10.1016/j.heliyon.2019.e01780

[35] Mehmood, K., Afzal, H., Majeed, A., & Latif, H. (2015). Contributions to the study of bi-lingual Roman Urdu SMS Spam Filtering. Software Engineering Conference (NSEC), 2015 National, 42–47.

[36] Mercado, R. N. M., Faustino, H., & Gloria, E. (2018). Automatic Cyberbullying Detection in Spanish-language Social Networks using Sentiment Analysis Techniques. International Journal of Advanced Computer Science and Applications, 9(7). https://doi.org/10.14569/IJACSA.2018.090733

[37] Mouheb, D., Abushamleh, M. H., Abushamleh, M. H., Al Aghbari, Z., & Kamel, I. (2019). Real-Time Detection of Cyberbullying in Arabic Twitter Streams. 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 1–5.

[38] Mouheb, D., Ismail, R., Al Qaraghuli, S., Al Aghbari, Z., & Kamel, I. (2018). Detection of Offensive Messages in Arabic Social Media Communications. 2018 International Conference on Innovations in Information Technology (IIT), 24–29.

[39] Nair, R. R., Mathew, J., Muraleedharan, V., & Kanmani, S. D. (2019). Study of Machine Learning Techniques for Sentiment Analysis. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 978–984.

[40] O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. Pediatrics, 127(4), 800–804.

[41] Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017). Detection of cyberbullying on social media messages in Turkish. 2017 International Conference on Computer Science and Engineering (UBMK), 366–370.

[42] Pawar, R., & Raje, R. R. (2019). Multilingual Cyberbullying Detection System. 2019 IEEE International Conference on Electro Information Technology (EIT), 040–044.

[43] Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., & Ungar, L. (2015). The role of personality, age, and gender in tweeting about mental illness. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 21–30.

[44] Rafae, A., Qayyum, A., Moeenuddin, M., Karim, A., Sajjad, H., & Kamiran, F. (2015). An unsupervised method for discovering lexical variations in roman urdu informal text. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 823–828.

[45] Raisi, E., & Huang, B. (2017). Cyberbullying detection with weakly supervised machine learning. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 409–416.

[46] Rajeshkanna, A., Preetha, V., & Arunesh, K. (2019). Experimental Analysis of Machine Learning Algorithms in Classification Task of Mobile Network Providers in Virudhunagar District. International Conference on E-Business and Telecommunications, 335–343.

[47] Riaz, K. (2008). Concept search in Urdu. Proceedings of the 2nd PhD Workshop on Information and Knowledge Management, 33–40.

[48] Sampasa-Kanyinga, H., Roumeliotis, P., & Xu, H. (2014). Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren. PloS One, 9(7), e102145.

[49] Sandhya, N. (2019). A Survey on Data Science Approach to Predict Mechanical Properties of Steel. International Conference on E-Business and Telecommunications, 501–511.

[50] Saravanaraj, A., Sheeba, J. I., & Devaneyan, S. P. (2016). Automatic Detection of Cyberbullying from Twitter. International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN, 2249–9555.

[51] Sheeba, J. I., & Devaneyan, S. P. (2017). Cyberbully Detection from Twitter using Classifiers. 9(8), 7.

[52] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. Australasian Joint Conference on Artificial Intelligence, 1015–1021.

[53] Vallabhaneni, S. R. (2019). Wiley CIA Exam Review 2019, Part 3: Business Knowledge for Internal AuditingElements. John Wiley & Sons.

[54] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS One, 13(10), e0203794.

[55] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., & Hoste, V. (2015). Automatic detection and prevention of cyberbullying. International Conference on Human and Social Analytics (HUSO 2015), 13–18.

[56] Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.

[57] Weir, G. R., Toolan, F., & Smeed, D. (2011). The threats of social networking: Old wine in new bottles? Information Security Technical Report, 16(2), 38–43.

[58] Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. Journal of School Violence, 14(1), 11–29.

[59] Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 656–666.

[60] Yang, C., Harkreader, R., Zhang, J., Shin, S., & Gu, G. (2012). Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter. Proceedings of the 21st International Conference on World Wide Web, 71–80.

[61] Yardi, S., Romero, D., & Schoenebeck, G. (2010). Detecting spam in a twitter network. First Monday, 15(1).

[62] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2, 1–7.

[63] Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3.