DEVELOPMENT OF COMPUTATIONAL LINGUISTIC RESOURCES FOR AUTOMATED DETECTION OF TEXTUAL CYBERBULLYING THREATS IN ROMAN URDU LANGUAGE

Amirita Dewani
Mehran University of Engineering & Technology, Jamshoro, Sindh, (Pakistan).
E-mail: amirita@faculty.muet.edu.pk ORCID: https://orcid.org/0000-0002-3816-3644

Mohsin Ali Memon
Mehran University of Engineering & Technology, Jamshoro, Sindh, (Pakistan).
E-mail: mohsin.memon@faculty.muet.edu.pk ORCID: https://orcid.org/0000-0003-2638-4252

Sania Bhatti
Mehran University of Engineering & Technology, Jamshoro, Sindh, (Pakistan).
E-mail: sania.bhatti@faculty.muet.edu.pk ORCID: https://orcid.org/0000-0002-0887-8083

Recepción: 20/04/2021 **Aceptación:** 10/06/2021 **Publicación:** 29/06/2021

Citación sugerida:
Dewani, A., Memon, M. A., y Bhatti, S. (2021). Development of computational linguistic resources for automated detection of textual cyberbullying threats in Roman Urdu language. *3C TIC. Cuadernos de desarrollo aplicados a las TIC, 10*(2), 101-121. https://doi.org/10.17993/3ctic.2021.102.101-121

ABSTRACT

Automatic Cyberbullying detection has remained very challenging task since social media content and conversations are usually posted in unstructured free-text form leaving behind the language norms. The major concern and gap in formulating cyberbullying detection strategies is scarcity of available linguistic resources typically for newly evolved languages. Roman Urdu has recently emerged and hence is a resource poor language. Urdu has been widely known as the national language of Pakistan. However, because of socio-cultural and multilingual aspects, Roman Urdu is used widely on the Internet by Asians and more specifically Pakistanis.

To fulfil the above stated gap, this research work presents guidelines for data annotation process and developed two linguistic resources: (i) Annotated corpus in Roman Urdu Language for cyberaggression and offensive language detection. The process of data annotation involved bilingual annotators instead of crowdsourcing. It has the benefit of correctly annotating instances that constitute clear cases of cyberbullying without compromising data quality. The developed corpus is highly balanced (with almost negligible skew) unlike most of the existing corpuses even in mature languages. (ii) Processing textual information for NLP tasks involves Stop-word elimination as a sub phase. Stop words carry least semantic information and increase feature space as compared to the other tokens and index terms in corpora. We have developed domain specific stop words for Roman Urdu Language considering all the lexical variants and typically in the context of aggression detection and collected data. The work has been carried out using python programming language and Pycharm IDE.

KEYWORDS

Linguistic Resources, Cyberaggression, Cyberbullying, Hate Speech Detection. Abusive Language Automated Detection.

https://doi.org/10.17993/3ctic.2021.102.101-121

1. INTRODUCTION

The rapid advancement in technology and compelling needs of users have made internet and typically SNS's an integral part of everyone's life, resulting in huge amount of user generated content aka Big social media data. Escalation in Social media has completely shifted the way in which people view, create or share information and ideas (Namdeo *et al.*, 2017).

Undeniably, Web 2.0 has a vital role in the communication, relationships, and collaboration in today's society. The communities belonging to different age groups (children, youngsters, and adults) interact with each other anytime, anywhere in diverse ways (e.g. via laptops, smartphones, tablets etc.) and using wide number of social networking platforms. Even though the perks and positive edges of digital communication are evident since most of the user's internet usage is harmless but the anonymity preservation and freedom of speech often makes young people to be offensive and vulnerable leading towards one of the alarming threat i-e cyberbullying/Cyberaggression or hate speech (Van Hee *et al.*, 2018). People, typically youngsters have reported life disturbing and annoying experiences thus drawing the attention of researchers/scholars and making cyberbullying and its automatic detection a growing community need and a promising area of Natural Language Processing (NLP) (Huang *et al.*, 2018).

Several studies contributed by different researchers are evident that computational formation of cyberaggression detection strategies is extremely challenging. One of the major challenges is posed by the scarcity of the required resources typically for newly emerged languages. Moreover, most of the datasets used for cyberbullying detection, even in mature languages, exhibit an extreme skew between hate speech and non-hate speech textual contents (Emmery *et al.*, 2020). This leads to formation of inappropriate strategies, unreliable predictive performance (specifically for the minority class) and more sensitivity towards classification errors.

With advent of Unicode encoding, Urdu language content, written using roman script, is escalating rapidly on social networking sites. Roman Urdu is a nonnormative language. The written script of

https://doi.org/10.17993/3ctic.2021.102.101-121

this language does not follow any rigid set of grammatical rules or standards of spellings. A survey statistics in (Shahroz *et al.*, 2020) affirms that about 300 million people are speaking Urdu language and approximately 11 million speakers are in Pakistan from which maximum users switched to Roman Urdu language for the textual communication, typically on social media (Shahroz *et al.*, 2020). It is linguistically rich and morphologically complex language (Mehmood *et al.*, 2020).

Urdu orthography (aka imla) bears a resemblance to Trukish, Arabic and Persian languages. Moreover, cursive Arabic and Nastaliques writing style is used (Syed *et al.*, 2010). Roman Urdu uses Roman script. An example instance of Roman Urdu script and its equivalent Urdu and English scripts are depicted in Figure 1.

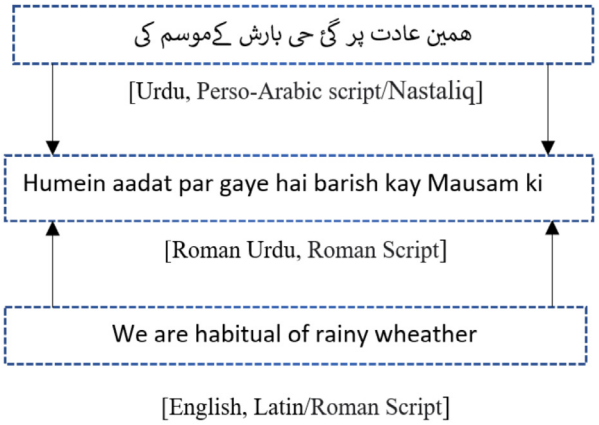


Figure 1. Script and Morphological variability in Roman Urdu, Urdu, and English Language.
Source: own elaboration.

Regardless of its huge prevalence worldwide (and more specifically in South Asia), Roman Urdu is an under-resource language. Linguistic resources for Asian languages are typically focused by some of the conferences and journals such as ACM Transactions on Asian and Low-Resource Language Information Processing ("ACM Transactions on Asian and Low-Resource Language Information Processing", n.d.), International Joint Conference of Natural Language Processing ("First Task for Automatic Cyberbullying Detection for the Polish Language | ACL Member Portal", 2019), Conference

https://doi.org/10.17993/3ctic.2021.102.101-121

of Central Asian Language and Linguistics ("Central Asian Languages and Lingusitics", n.d.) etc. for supporting vast number of NLP tasks related to phonology, morphology, name entity recognition (NER), language parsing and word segmentation.

To support the development of NLP applications for Roman Urdu typically in the field of cyberaggression and hate speech, this paper presents annotation guidelines, the first-ever highly balanced Roman Urdu dataset and development of domain specific stop words using python language.

Rest of this paper is organized as follows: Cyberaggression and existing resources are conferred in section II. Section III puts light on Data extraction from Twitter social media platform. Data annotation guidelines preparation and kappas weighing scheme are given in section VI & V respectively. Section VI discusses Stop word development. Finally, Section VII conclude the research work.

2. RELATED WORK

Even though the researchers have widely used Natural Language Processing (NLP) and realized Machine Learning (ML) techniques to uncover solutions for variety of tasks based on unstructured text data (e.g. topic identification, opinion mining, document summarization, text translation etc.), but it's applicability for resolving automatic detection of cyber-crime related problems is relatively new and has encountered so many challenges (Rosa *et al.*, 2019).

The availability of appropriate data, huge data skew because of natural uneven distribution of hate speech content on social media and NLP resources scarcity represents one amongst many significant issues in research on cyberbullying detection (Mahlangu *et al.*, 2018; Gencoglu, 2020). A handful of studies are contributed by scholars to develop resources and cyberbullying detection strategies in different languages worldwide. Most studies have hateful instances ranging from 2 to 5% (Emmery *et al.*, 2020).

The study by Sprugnoli *et al.* (2018), developed a WhatsApp dataset from WhatsApp chats to study offensive language among Italian students. They also presented annotation scheme and user roles.

https://doi.org/10.17993/3ctic.2021.102.101-121

Research work accomplished in Fersini *et al.* (2018) collected misogynous and hateful tweets data using a combined approach. They monitored prospective victims of hate accounts, downloading the history of identified haters and filtered twitter stream contents via keywords.

The study conducted in Fišer *et al.* (2017) extracted data from an online platform that collects impulsive reports by internet users of any material having Child sex abuse; a special category of cyber-aggressiveness, to develop a corpus. The validation of corpus by experts revealed that only 3% was illegal content and more than 40% in non-disturbing content. Indonesian language hateful corpus was contributed in Ibrohim and Budi (2018). The research work used twitter platform, crowdsourcing annotation, and a multi-level scheme to identify Hate speech and non-hate speech categories along with their intensity levels. Work carried out in Bohra *et al.* (2018) presents a dataset comprising of Hindi-English code-mixed data. The tweets are annotated with the language at word level and the class they belong to (Hate Speech or Normal Speech).

The study in Van Bruwaene *et al.* (2020) formed a dataset using multiple platforms in English language from SafeToNet's VISR-branded child safety app for adolescents. In collaboration with expert annotators, they utilized crowd sourcing and machine learning techniques to enlarge the corpus and handle skew in iterative manner. The work by Özel *et al.* (2017) is the first study performed in Turkish Language. The research has contributed corpus in Turkish language prepared using Instagram and twitter social media platforms. Experimentation is also conducted using machine learning techniques.

Undeniably, English is the de facto common language among researchers at international level, hence greater number of computational resources, as highlighted by a review study (Poletto *et al.*, 2020) are English corpora and datasets. Nevertheless, several other languages are represented too, and this certainly is immensely significant for international community that seeks to address a worldwide social issue of cyberbullying and hate speech spread in many languages.

https://doi.org/10.17993/3ctic.2021.102.101-121

Roman Urdu has become a contemporary trend these days as a language of communication for Pakistani or more generally Asian youth. To the best of knowledge, this is the first ever study that has developed computational linguistic resource of Roman Urdu for Cyberaggression. This research study presents our approach for collecting and annotating social media data to develop a cyberbullying corpus in Roman Urdu language and domain specific stop words. The extraction of data was a multi-phase process to ensure high quality data with minimum skew. It encompasses vast range of content inciting hatred. The content is also based on wide bullying tactics like race, ethnic origin, religious affiliation, sexual orientation, caste, gender, identity and serious disease or disability Intelligence. Since a natural distribution of social media data is heavily skewed which results in a scarcity of bullying instances to be used in training, So we did extraction in phases. Moreover, this work used different weighing schemes for automatic identification and bilingual expert annotators manual input to develop stop words related to cyberbullying detection problem in Roman Urdu.

3. METHODOLOGY

3.1. DATA EXTRACTION

Twitter is one of the most popular microblogging service having 316 million monthly active users. As compared to other social media platforms, twitter has attracted more to the academic researchers as it makes its data available for research purposes via Application Programming Interface (API) (Ahmed *et al.*, 2017). To develop cyberbullying corpus, data was scrapped from twitter using python language, tweepy and twitter streaming API in multiple phases over the duration of 3 months as depicted in Figure 2. The reasons were twofold: (i) Restrictions on data access imposed on standard API (ii) The natural distribution of content is highly skewed.

https://doi.org/10.17993/3ctic.2021.102.101-121

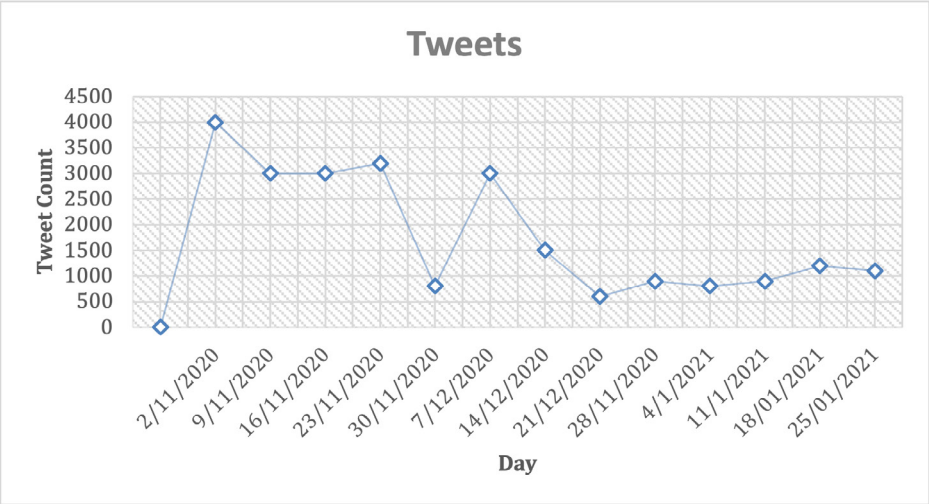


Figure 2. Data Extraction- Tweet Count.
Source: own elaboration.

Currently, no language code is available for Roman Urdu in API, So the queries for data collection were formed based on geo-location information; taking coordinates from google maps of the areas in Pakistan where high saturation of Roman Urdu content was expected. Secondly, we extracted tweets based on insulting seeds or curse words typically used in Roman Urdu language for bullying. Thirdly we used hash values for aggression and trash talking on recent topics from the regions in Pakistan. Substantial number of tweets were in English Language and such content was filtered out leaving behind 3K tweets. In order to retain writing patterns of Roman Urdu users on social media, data with inherent English words (*such as batting, topic, character, bowling, follow, design, ok, yes, no, music, video, free, hope, player, code, development, etc.*) was preserved. Some examples of such data instances are depicted in Figure 3.

https://doi.org/10.17993/3ctic.2021.102.101-121

2021-01-02 15:03:03 ,b' Yakeen karen …. Follow back milnay kay bad unfollow karny waly log kisi ky sagge nahe ho sakty …, "b'Karachi, Pakistan'"
2020-12-15 18:44:06,b' Acha atleast tell me boards k papers postpone hone ki koi probability hai?', "b' Faisalabad, Pakistan'"

Figure 3. Natural writing patterns of users on social media in Roman Urdu Language.
Source: own elaboration.

3.2. DATA ANNOTATION PROCESS

Data annotation is indeed a Human Intelligence Task (HIT). Undeniably, crowdsourcing has obvious organizational advantages, especially for a time consuming task as the annotation of textual data, but annotation quality might get compromised from employing non-expert annotators typically for recently evolved languages and challenging task like cyberaggression (Schmidt & Wiegand, 2017). Moreover, many studies have uncovered that a non-trivial percentage of the data collected on MTurk is "doubtous", annotated either by "non-respondents" (bots instead of humans) or non-serious respondents (Dreyfuss *et al.*, 2018; Ahler *et al.*, 2019).

Instead of crowdsourcing, data was annotated by linguistic experts having bilingual expertise (having good knowledge of Nastaliq scripting and Roman Urdu patterns). Annotators were provided with guidelines on how to label social media documents for bullying. The main task of each annotator was to label each sample with one of three possible labels:

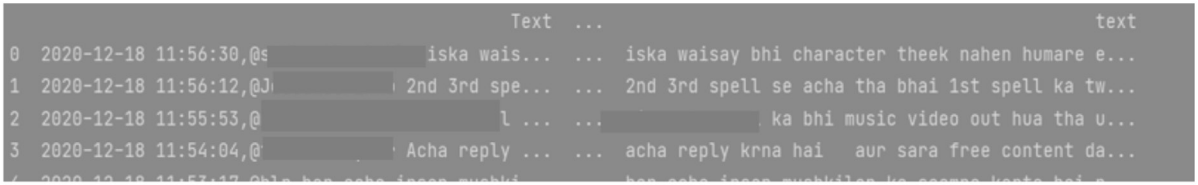
0 – text certainly does not contain any form of online violence, hate speech or abusive language.

1 – text certainly contains any form of online violence, hate speech or abusive language

2 – Indeterminate case(doubtful) when text cannot be identified with good certainty to either contain or do not contain any form of online violence, hate speech or abusive language.

https://doi.org/10.17993/3ctic.2021.102.101-121

To provide the annotators with some context and preserve original writing intentions, all posts were presented in their original form i-e before applying major text preprocessing techniques, wherever possible. Social Media data is considered as Microtext (Mehmood *et al.*, 2020). Microtext is extremely noisy. Deep comprehension of microtext is immensely important for effective understanding and further processing. Hence preliminary Text preprocessing was applied to handle Unicodes, punctuations, hashtags, emojis, URLs, case conversion, date and time data, insignificant string literals, and @ user mentions. The preprocessing results are given in Figure 4.



		Text	...		text
0	2020-12-18 11:56:30,@s	iska wais...	...	iska waisay bhi character theek nahen humare e...	
1	2020-12-18 11:56:12,@J	2nd 3rd spe...	...	2nd 3rd spell se acha tha bhai 1st spell ka tw...	
2	2020-12-18 11:55:53,@	l ...	...	ka bhi music video out hua tha u...	
3	2020-12-18 11:54:04,@	Acha reply ...	...	acha reply krna hai aur sara free content da...	

Figure 4. Preliminary Text preprocessing on Roman Urdu Data.
Source: own elaboration.

To avoid any incorrect annotation, the linguistic experts were encouraged to use "2" whenever they have doubts if an instance contains any form of aggressive speech or not. The example of such doubtful instances is given in Figure 5.

2020-12-18 09:42:05,b' Inse acha to aj kal gali cricket match me ache catch hote hai', b"
2020-12-18 01:31:55,"b""iska jab dosri larki kay sath affair tha he to sara ki zinadgi tabah ki
2020-12-18 09:25:42,b' Ziyada Free Na Ho dono main interest the hai,b"

Figure 5. Examples of instance might be tagged as 2 by annotators.
Source: own elaboration.

Because of the huge size of dataset, the whole annotation process was split into multiple phases. The entire annotation process was conducted by linguistic experts as per developed guidelines and categories.

https://doi.org/10.17993/3ctic.2021.102.101-121

Based on these phases, the final labels were determined using weighting scheme described in section 5. The main concern of annotators was to analyze which types of phenomena in Roman Urdu can be considered as cyberbullying/hate speech (e.g. attack on personality characteristics, threats, curse, blackmails) or profanity/ abusive language and aggressiveness that might harm an individual's physical or mental health, lower self-esteem or hurt feelings. The sample data instances for each category are given in Table 1.

Table 1. Cyberbullying categories and Roman Urdu Instances.

S. No	Category	Example Instances
1	Threat or blackmail	beta Himat hai na to mujhse milnay aajana batata hon phr' Main tumhaun choron ga nahe tum bas dekhti jao.
2	Racism	Bhai extra bounce jidr ho hum Acha ni krte, remember World Cup main West Indies shediyon ne kiya kia tha Humare saath African jaise hogaye ho shakal dekho apni, jutt zameendaar log hain ye to phir dimag bhi to utna chalay ga
3	Insult	Ji uncle ap chaly jayein please twitter suit nahe krta ap pay Subha subha mazaq acha nai lgta shaista ko bolo apna talak karwaya ab doosron ka karwao:/'
4	Sexual Talk	kapray kam hote ja rahe hain, Oh babs, shayad sex education topic nahe parha main parha doon baby
5	Curse or exclusion	yes, you are the only one jo itna bura hy. dua kar raha hoon k tum mar jao. Allah karay tum hamesha aziyat main raho takleef main raho kabhi khush na raho
6	Defamation	don't expect ye jhooti aurat acha acha sach bolegi Hhhhhhhhh acha mazak hai lanat ho tm par aur tumhary channel pr'
7	Encouragements to the harasser	Sahi kah raha hai tum mar kiyon neh jati itni badnami kafi neh hai kiya
8	Personality characteristics	Ary ry ry tum to pooray aalo ban gaye ho ek hota hai kaala aur phir aata hai daambar ka tukra

Source: own elaboration.

3.3. WEIGHING SCHEME

Since the process of annotation is a subjective phenomenon, hence different annotators might have a bit different judgment for the same textual comments. To eliminate dubious data and to ensure data quality

https://doi.org/10.17993/3ctic.2021.102.101-121

and for further scrutiny and validation of dataset, Inter-Rater Reliability (IRR/IRA) was estimated using Cohen's kappa coefficient (κ) (definition 5.1). for measurement of inter-annotator agreement, Cohen's kappa coefficient has been accepted as the de facto standard. Empirically, based on kappa score the threshold value or cut off value is set. Usually, kappa score of 0.67 is used as a cutoff in computational linguistics (Wang *et al.*, 2019).

3.3.1. DEFINITION AND COMPUTATION OF COHEN' S KAPPA

Cohen's kappa is a function of *po*, the relative observed agreement, and *pe*, the expected hypothetical agreement by chance. Mathematically it can be stated as in equation 1.

$$\kappa = \frac{Po - Pe}{1 - Pe} \quad \text{..... (1)}$$

Where *p*0 i-e observed agreement can be computed using Equation (2).

$$p0 = \frac{\text{count of agreements}}{\text{ount of agreements+count of disagreements}} \quad \text{..... (2)}$$

and *pe* i-e the expected agreement by chance can be computed using Equation (3).

$$pe = \frac{1}{N^2} \sum k \; nk1 \; nk2 \quad \text{..... (3)}$$

Where N indicates the number of items, *k* is the number of categories, *nk*1 is the number of times rater 1 selected category *k*, and *nk*2 is the number of times rater 2 selected category *k*.

The kappa score was computed using Statistical Software (n.d.). The standard error and 95% confidence interval are calculated according to Fleiss *et al.* (2013). The weighted kappa score as shown in Table 2 is significantly higher than the cutoff value.

Table 2. Weighted Kappa Score.

Weighted Kappa	0.81818
Standard Error	0.12730
95% CI	0.56868 to 1.00000

Source: own elaboration.

https://doi.org/10.17993/3ctic.2021.102.101-121

4. RESULTS

4.1. IDENTIFICATION OF DOMAIN SPECIFIC STOP WORDS

In this era of big data and information retrieval, process optimization for Text and Data Analytic systems becomes immensely significant. Therefore, to achieve accuracy, identification & filtration of terms with minimal or no semantic meaning is significant. Stop words has been developed in so many languages like English, Italian, Chinese, Arabic, Punjabi, Hindi, etc. (Kaur & Saini, 2015, 2016; Hao & Hao, 2008; Alajmi *et al.*, 2012) and are also part of NLTK, spaCy, and gensim.

Stop words not only vary from corpora to corpora, language to language, they also vary from one problem domain to another. For example, in a corpus of news articles, comprising of crucial information that is time-sensitive and location-sensitive, eliminating terms like "here", "today", etc. would affect results of related NLP application. This is because news articles link and relate current event to the similar events that had happened in the past or on another location.

In this work, we have identified stop-words from developed corpora, specific to the domain of cyberbullying and hate speech using statistical methods and human evaluation i-e direct Term Frequency (TF), Inverse-Document Frequency (IDF), Term-Frequency-Inverse-Document-Frequency (TFIDF) weighting model and human evaluation by bilingual experts. The methods are described below.

Let tf(t,d) denotes direct frequency of term t in document d. Mathematically, it can be defined as:

$$tf\ (t,\ d) = fd(t) \ / \ maxw€d\ fd(w)$$

where fd(w) indicates the total number of words in a document. The metric term frequency highlights commonality of a term within a collection of documents. Semantically least significant terms are expected to have high term frequency. 30 samples, based on term frequencies are depicted in Figure 6.

https://doi.org/10.17993/3ctic.2021.102.101-121

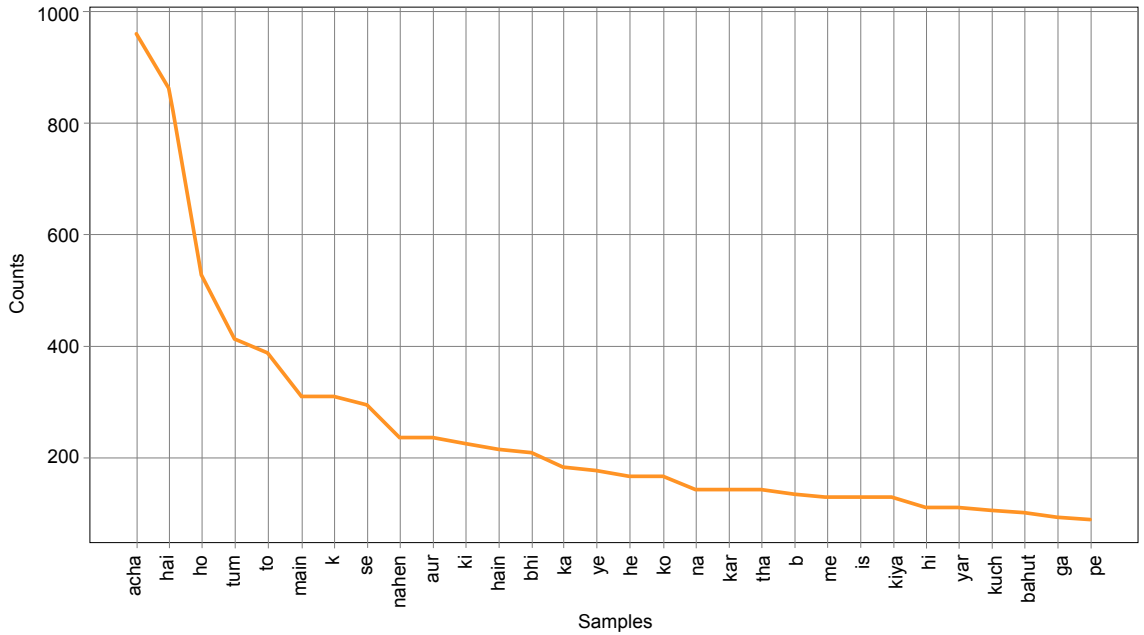


Figure 6. Samples based on term frequencies (n=30).
Source: own elaboration.

Inverse document frequency, idf can be defined mathematically as:

$$Idf\ (t,D) = ln\ [(/D/)/\ \{/d \in D: t \in D/\}]$$

Where, D is the collection of documents in the corpus.

This metric is very significant since it penalizes the more frequently occurring terms and favors the ones occurring in a few documents only. The lower bound of this metric is 0 and refers to the terms that appear in every single document in the corpus 40. Feature names sorted by their idf weights in ascending order are given in Figure 7.

https://doi.org/10.17993/3ctic.2021.102.101-121

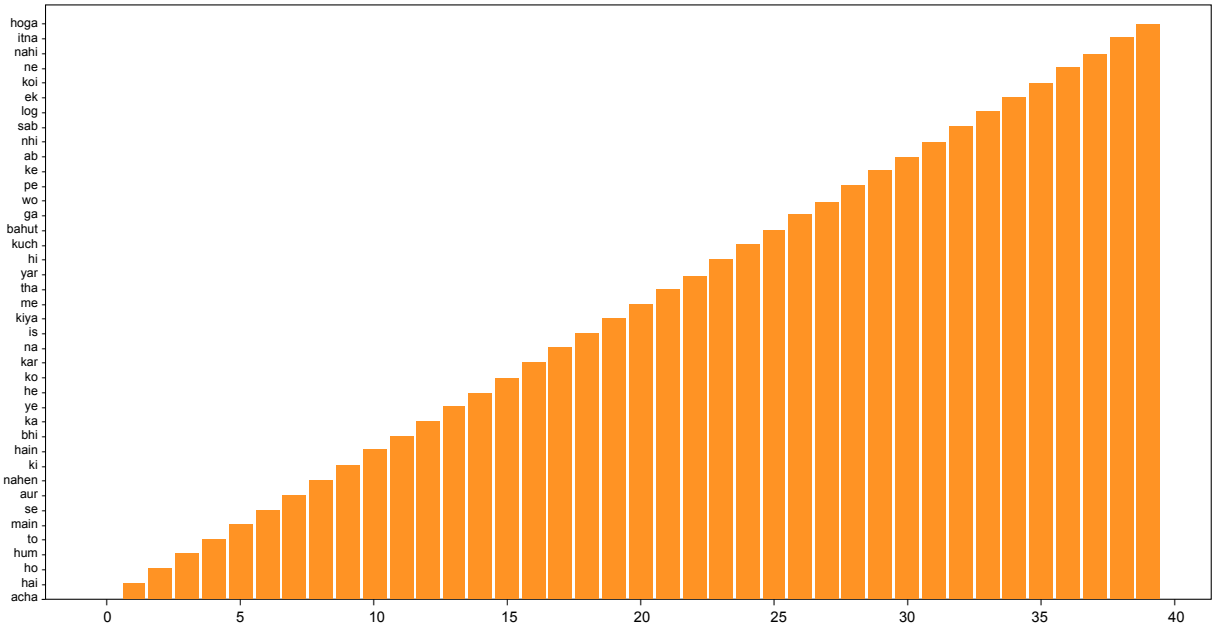


Figure 7. Feature names by idf weights (Sorted in ascending order).
Source: own elaboration.

```
['aap' 'ab' 'abhi' 'acha' 'ache' 'achi' 'agar' 'aise' 'ajeeb' 'allah'
 'and' 'ap' 'apko' 'apna' 'apni' 'aur' 'aye' 'baad' 'baat' 'bahut' 'ban'
 'bana' 'bas' 'bat' 'bhai' 'bhi' 'bht' 'bilkul' 'birthday' 'but' 'chaiye'
 'chal' 'chor' 'de' 'dekh' 'dekha' 'dekho' 'dil' 'dimag' 'din' 'diya' 'do'
 'dono' 'dont' 'dost' 'ek' 'ga' 'gay' 'gaya' 'gaye' 'ghar' 'gi' 'good'
 'ha' 'hai' 'hain' 'hamesha' 'har' 'he' 'hi' 'ho' 'hoga' 'hogaye' 'hon'
 'hona' 'hoon' 'hota' 'hoti' 'hu' 'hua' 'hum' 'hun' 'hy' 'in' 'insan' 'is'
 'itna' 'itne' 'itni' 'ja' 'jab' 'jaisay' 'jao' 'jaye' 'jo' 'ka' 'kaam'
 'kabhi' 'kahin' 'kam' 'kar' 'karay' 'karen' 'karna' 'karo' 'karta'
 'karte' 'kay' 'ke' 'khud' 'khush' 'ki' 'kia' 'kisi' 'kiya' 'kiyon' 'ko'
 'koi' 'kr' 'krna' 'kro' 'kuch' 'kya' 'lag' 'laga' 'lagta' 'larki' 'le'
 'liye' 'lo' 'log' 'lolz' 'love' 'mai' 'main' 'maine' 'mar' 'mat' 'me'
 'mein' 'mera' 'meray' 'mere' 'mujhay' 'mujhe' 'my' 'na' 'nahe' 'nahen'
 'nahi' 'ne' 'nhi' 'of' 'oh' 'or' 'par' 'pass' 'pata' 'pe' 'pehle' 'phir'
 'phr' 'please' 'raha' 'rahay' 'rahe' 'raho' 'rha' 'sab' 'sachi' 'sahi'
 'sath' 'say' 'se' 'shakal' 'sirf' 'so' 'sorry' 'sy' 'tak' 'tarah' 'tha'
 'the' 'theek' 'thi' 'time' 'to' 'toh' 'tou' 'tu' 'tum' 'tumhain'
 'tumhara' 'tumhare' 'tweet' 'us' 'use' 'wah' 'wala' 'warna' 'wo' 'woh'
 'ya' 'yaar' 'yahan' 'yar' 'ye' 'yeh' 'you' 'ziada']
```

Figure 8. Terms sorted by low Tfidf scores (max-features= 200).
Source: own elaboration.

https://doi.org/10.17993/3ctic.2021.102.101-121

a b

Figure 9. Cloud Representation of Corpus. (a) Prior to removal of stop words. (b) After removal of stop words.
Source: own elaboration.

7. CONCLUSIONS

Social networking sites has become a communal breeding ground for youth to aggress one another. During the pandemic, the traffic in cyberspace increased significantly. Recent studies and published news reports highlight that there is great surge in the number of cyberbullying and harassment cases during the pandemic. This paper made novel contributions and achieved important landmark in regard of NLP on Roman Urdu language which has been embraced recently on social media by youth. The corpus was developed for cyberbullying and hate speech by collecting data for over 3 months to have high quality data with less skew. A well-defined set of data annotation guidelines were prepared and provided to experts for annotation. Since the process of annotation was a subjective phenomenon, hence, for further scrutiny and validation of corpus, Inter-Rater Reliability (IRR) based on Cohen's kappa coefficient(κ) was identified using statistical software Medcalc. Finally, the work rigorously developed an automatic stop word identification strategy using statistical methods and weighing model. Moreover, manual input

https://doi.org/10.17993/3ctic.2021.102.101-121

of linguistic experts was also taken to form comprehensive list specific to cyberbullying corpus. In future we plan to conduct experiments using deep learning approaches and hand engineered feature sets. As the field of cyberaggression and cyberbullying is at its emergence so this research would greatly benefit NLP and machine learning research community.

ACKNOWLEDGEMENTS

This research has been performed at Institute of information Technology, Mehran University of Engineering and Technology, Pakistan and is Funded under MUET funds for postgraduate students. We would like to thank bilingual experts Ms. Irum Parvaiz (Lecturer Dawood University of Engineering & Technology) and Mariyam Bughio (Assistant Professor, Zubaida Degree Girls college) for huge support through data annotation process.

REFERENCES

ACM Transactions on Asian and Low-Resource Language Information Processing. (n.d.). https://dl.acm.org/journal/tallip

Ahler, D. J., Roush, C. E., & Sood, G. (2019). *The micro-task market for lemons: Data quality on Amazon's Mechanical Turk.* Meeting of the Midwest Political Science Association.

Ahmed, W., Bath, P. A., & Demartini, G. (2017). *Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In The Ethics of Online Research* (pp. 79–107). Emerald Publishing Limited.

Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Toward an ARABIC stop-words list generation. *International Journal of Computer Applications, 46*(8), 8–13. https://www.researchgate.net/publication/306364790_Toward_an_ARABIC_Stop-Words_List_Generation

https://doi.org/10.17993/3ctic.2021.102.101-121

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media,* 36–41.

Central Asian Languages and Lingusitics. (n.d.). https://concall.indiana.edu/papers.html

Dreyfuss, E., Barrett, B., & Newman, L. H. (2018). *A bot panic hits Amazon's Mechanical Turk. Wired.* https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., & Daelemans, W. (2020). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*, 1–37. https://www.researchgate.net/publication/336869277_Current_Limitations_in_Cyberbullying_Detection_on_Evaluation_Criteria_Reproducibility_and_Data_Scarcity

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian, 12,* 59. http://personales.upv.es/prosso/resources/FersiniEtAl_Evalita18.pdf

First Task for Automatic Cyberbullying Detection for the Polish Language | ACL Member Portal. (2019). https://www.aclweb.org/portal/content/deadline-extension-first-task-automatic-cyberbullying-detection-polish-language

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. *Proceedings of the First Workshop on Abusive Language Online,* 46–51. https://www.aclweb.org/anthology/W17-3007/

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

https://doi.org/10.17993/3ctic.2021.102.101-121

Gencoglu, O. (2020). *Cyberbullying detection with fairness constraints.* IEEE Internet Computing.

Hao, L., & Hao, L. (2008). Automatic identification of stop words in chinese text classification. In *2008 International Conference on Computer Science and Software Engineering, 1,* 718–722. https://www.semanticscholar.org/paper/Automatic-Identification-of-Stop-Words-in-Chinese-Hao-Hao/cd7d1fdf3b3eec1ebe9378065279442bacbf6fa5

Huang, Q., Inkpen, D., Zhang, J., & Van Bruwaene, D. (2018). Cyberbullying Intervention Interface Based on Convolutional Neural Networks. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, 42.* https://www.aclweb.org/anthology/W18-4405/

Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science, 135,* 222–229. https://doi.org/10.1016/j.procs.2018.08.169

Kaur, J., & Saini, J. (2016). Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. *International Journal of Computer Applications, 150,* 15–17. https://doi.org/10.5120/ijca2016911462

Kaur, J., & Saini, J. R. (2015). A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language. *Annals of Health and Health Sciences-Indian Journals, 5*(2), 114-120. https://doi.org/10.5958/2249-3220.2015.00015.4

Mahlangu, T., Tu, C., & Owolawi, P. (2018). A review of automated detection methods for cyberbullying. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC),* 1–5. https://www.semanticscholar.org/paper/A-Review-of-Automated-Detection-Methods-for-Mahlangu-Tu/fc86a8bab87e3cdb6cc1f35bf38094aa373d88de

https://doi.org/10.17993/3ctic.2021.102.101-121

Mehmood, F., Ghani, M. U., Ibrahim, M. A., Shahzadi, R., Mahmood, W., & Asim, M. N. (2020). A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access, 8,* 192740–192759. https://ieeexplore.ieee.org/document/9223633

Namdeo, P., Pateriya, R. K., & Shrivastava, S. (2017). A Review of Cyber bullying Detection in Social Networking. *Proceedings of ICICCT,* 162–170.

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

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation,* 1–47. https://doi.org/10.1007/s10579-020-09502-8

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. M. V., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior, 93,* 333–345. https://doi.org/10.1016/j.chb.2018.12.021

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://www.aclweb.org/anthology/W17-1101/

Shahroz, M., Mushtaq, M. F., Mehmood, A., Ullah, S., & Choi, G. S. (2020). RUTUT: Roman Urdu to Urdu Translator Based on Character Substitution Rules and Unicode Mapping. *IEEE Access, 8*, 189823–189841.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., & Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 51–59.

https://doi.org/10.17993/3ctic.2021.102.101-121

Statistical Software. (n.d.). https://www.medcalc.org/

Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2010). Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. In Sidorov, G., Hernández, A., Reyes, C. A. (eds) *Advances in Artificial Intelligence.* MICAI 2010. Lecture Notes in Computer Science, vol 6437. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16761-4_4

Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation, 54*(4), 851–874. https://doi.org/10.1007/s10579-020-09488-3

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. https://doi.org/10.1371/journal.pone.0203794

Wang, J., Yang, Y., & Xia, B. (2019). A Simplified Cohen's Kappa for Use in Binary Classification Data Annotation Tasks. *IEEE Access, 7,* 164386–164397.