IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Un-Compromised Credibility: Social Media based Multi-Class Hate Speech Classification for Text

KHUBAIB AHMED QURESHI[1], MUHAMMAD SABIH[2]
[1]DHA Suffa University, Karachi, Pakistan (e-mail: k.ahmed@dsu.edu.pk)
[2]DHA Suffa University, Karachi, Pakistan (e-mail:m.sabih@dsu.edu.pk)
Corresponding author: Khubaib Ahmed Qureshi (e-mail: k.ahmed@dsu.edu.pk).

ABSTRACT There is an enormous growth of social media which fully promotes freedom of expression through its anonymity feature. Freedom of expression is a human right but hate speech towards a person or group based on race, caste, religion, ethnic or national origin, sex, disability, gender identity, etc. is an abuse of this sovereignty. It seriously promotes violence or hate crimes and creates an imbalance in society by damaging peace, credibility, and human rights, etc. Detecting hate speech in social media discourse is quite essential but a complex task. There are different challenges related to appropriate and social media-specific dataset availability and its high-performing supervised classifier for text-based hate speech detection. These issues are addressed in this study, which includes the availability of social media-specific broad and balanced dataset, with multi-class labels and its respective automatic classifier, a dataset with language subtleties, dataset labeled under a comprehensive definition and well-defined rules, dataset labeled with the strong agreement of annotators, etc. Addressing different categories of hate separately, this paper aims to accurately predict their different forms, by exploring a group of text mining features. Two distinct groups of features are explored for problem suitability. These are baseline features and self-discovered/new features. Baseline features include the most commonly used effective features of related studies. Exploration found a few of them, like character and word n-grams, dependency tuples, sentiment scores, and count of 1st, 2nd person pronouns are more efficient than others. Due to the application of latent semantic analysis (LSA) for dimensionality reduction, this problem is benefited from the utilization of many complex and non-linear models and CAT Boost performed best. The proposed model is compared with related studies in addition to system baseline models. The results produced by the proposed model were much appreciating.

INDEX TERMS Machine Learning, Multi-Class Hate Speech, Natural Language Processing, Hate Speech Classification, Social Media Microblogs, , Multi-Class Hate Speech Dataset, Twitter Hate Speech, Text Mining, Features Exploration

I. INTRODUCTION

SOCIAL media is massively used for different forms of content sharing. People extensively use social media to share their opinions and insights. Despite that social media is extremely fast, open, free, and easy to access, due to its explosive spreading nature it is quite vulnerable too. It turns into a medium for wrongdoers to spread different forms of hate or prejudice communication towards another group. Hate speech is essentially a discourse that might be extremely harmful to the feelings of a person or group and may contribute towards brutality or insensitivity which shows irrational and inhuman behavior. Growth of online social media has also increased hate speech which is a crime. Hate speech and hate crimes are connected [2], it could also be seen that hate crimes are getting increased [1]. The problem of hate speech is getting increased popularity, therefore many initiatives are also conducted at the government level, e.g.: the Council of Europe executed the movement of No Hate Speech [10], legislation has also been made to eliminate its proliferation, named EU Hate speech code of conduct [6], which must be signed and implemented by all social media services within 24 hours. In this regard, Twitter was also accused by EU regulators of not being good regarding hate speech removal from their platform [13].

Hate speech detection is a challenging problem. There are disagreements in its definition, which make identification and annotations of hate speech more difficult and confusing from free expression [39]. Different aspects of definition from varying category of sources could be seen from, Twitter [3], YouTube [4], Facebook [5], International minorities associations ILGA [2], EU Commission's Code of Conduct [6], Encyclopedia of the American Constitution [7], American Bar Association [8], Davidson et al. [21], and finally, Fortuna et al. [9] where an effort has been made to explore many subtle aspects of hate definition and therefore same is followed for our data annotation in the study. In continuation to hate speech definition understanding which is a complex phenomenon, it is worth mentioning that there are many closely related concepts that are not assumed as hate speech, though few of them are confusingly considered hate speech in studies [20]. Those related concepts are: Hate, Cyberbullying, Discrimination, Flaming, Harassment, Abusive Language, Profanity, Toxic Language or comment, Extremism, Radicalization [9]. In contrast to related concepts, it has also been identified in studies [11] that there are different types of hate speech as well concerning its categories or targets on social media, e.g.: race, religion, ethnicity, gender, class, sexual orientation, behavior, physical, disability, and other (i.e. drunk, shallow people). Automatic hate speech detection is technically a difficult task, considering some challenging aspects of language subtleties among many others are, hate speech may not have any aggressive, offensive, profane, or derogatory terms but still categorized as hate speech and same is true for vice versa [21]. Similarly, all hate is not necessarily considered as hate speech [9]. Another challenge to hate speech detection is limited data availability over social media due to the enforcement of the hate speech code of conduct. Likewise, those seeking to spread such contents in presence of these legislations, are actively trying to find alternatives to circumvent complex measures put in place, which become more challenging for automatic detection [12].

The majority of these challenges discussed, are basically related to the quality of the dataset, which will all be addressed through quality-based strong datasets compilation, within this study. The next challenge which is also targeted in this study is to explore and identify the best set of features and then develop an appropriate classifier for hate speech detection. Considering dataset compilation, the highest categories of hate crimes reported by the FBI, are based on race, ethnicity, religion, and sexual orientation [1]. Therefore all these categories are primarily selected for datasets compilation (see these categories in table 3). In a variety of data science applications and detailed analysis, a fine-grained level of hate discourse is expected rather than simple hate speech classification, and there may be multiple hate targets expected in single hate speech discourse. There is no such study found to the best of our knowledge and this gap is being filled through our study. Regarding the selection of social media platforms, Twitter is accused by Europeans that they are extremely poor in hate speech removal from their platform, therefore it is targeted for data collection [13].

There are many challenges highlighted which are addressed through the contributions of the study. These contributions are following:

1. Availability of standard and appropriate dataset could guarantee the effective and high performing hate speech detection system. Therefore compilation of high quality, social media-specific, broad, and balanced datasets are achieved in the study. They could be used in many research studies and applications. They are named as 'Binary Classified Multi-Category Hate Speech Datasets', which include the following. The importance of datasets construction and how the challenges are addressed is discussed in section III.

a) 10 hate speech categories based datasets each with binary classes.
b) A combined dataset with multi-class hate speech labels.
c) Datasets with language subtleties.
d) Datasets labeled under comprehensive, clear definition and well-defined rules/ guidelines of hate speech.
e) Datasets with the strong agreement of annotators.

2. The next contribution of the study is to explore and identify the best set of text mining features. These features are extracted from related studies in addition to our own proposed features. Based on the feature analysis and identification an appropriate classifier for hate speech detection is developed. This include the following.

a) In addition to our own proposed features, most commonly used and effective text mining based features reported in studies are extracted for detailed exploration. These commonly used and effective features are treated as baseline features in our explorational study.
b) These set of IR, NLP, and Text Mining based features are completely explored and presented the analysis for the researchers of the field.
c) Identify the best set of features for problem suitability.
d) Experiments conducted using different classes of Machine Learning models. It includes linear, non-linear, tree-based, non-parametric, large margin classifiers, and Ensemble (boosting, bagging) models. Finally found that non-linear, tree-based, boosting models were best performing for the problem solution.
e) Proposed model's performance is compared with our system's baseline and other related studies.

The remaining of the paper is organized as follows: related work is briefly discussed in section II. The process of dataset construction with the annotation method is explained in section III. Complete experimental setup including data preprocessing, potential features, exploration of models, features, and best model and features selection are discussed in section IV. Results and Discussion is presented in sectionV.

Future directions are presented in the section VII followed by the conclusion in the sectionVI.High-level system framework is presented in figure 1 for better understanding.

II. RELATED WORK

Using Text Mining (TM), Information Retrieval (IR), or Natural Language Processing (NLP) for hate speech identification is considered much effective [9], [14] as compared to Keyword-based, Rule-based/Association Rule Mining [19], Source metadata/Social Network Analysis based approaches [12], [23]. Therefore most of the related work includes NLP, IR, and TM-based approaches. Authors of the research study [24] collected data of 16K tweets (WaseemA) using the list of terms and annotate them as sexist and racist for their supervised classification. Different word level n-grams of 1-4, and character level n-gram are used as features. In addition to n-grams, the user's gender, location, and description along with totals and averages of tweet length, word length, and user description length are also used. Character n-grams of length four, with other features, were found much better than word n-grams. Logistic Regression classifier performed as the best model in their experiments. An extended study [34] was also performed with few additional features (POS Tags and Skip grams) and an extended dataset (WaseemB), to explore the annotator's influence over classification performance. They found that experts annotated data outperformed. It is explored in an interesting study [21] that the identification of hate speech and offensive language is also a challenge. HatebaseTwitter dataset was constructed using Hatebase lexicon's [36] terms for tweets fetching and labeled as hate, offensive, and neither. Features used for this supervised classification task were: word-level uni, bi, and tri-grams, with TF-IDF weights, and part-of-speech tags, two different text readability scores, sentiment score, with social network features like: count of hashtags, mentions, retweets, and URLs. Length of the tweet, no of characters, words, and syllables are also used as features. These features are used for training and the best results were produced by the SVM classifier. Analysis of the results achieved by the mentioned study shows that homophobic and racist tweets are mostly identified as hate speech and sexist tweets are more likely to be classified as offensive. It is based on observation only since no formal or agreed-upon definition explicitly distinguishes hate speech from offensive language. It is a consensus that "hate speech is any expression targeted at disadvantaged groups that potentially incite violence or social disorder [38]". Another examination of methods [26] for achieving similar objectives of classifying tweets into hate, abusive, and neither. It involved the use of character n-grams (2-8), word n-grams (1-3), and word skip-grams (1-,2-,3- bi-grams) as features to a multi-class SVM model. It shows that the use of character 4-grams helps in accomplishing the best 78% accuracy on Davidson et al. dataset HatebaseTwitter [21], while being easier and more effectively interpretable choices than neural techniques [25]. Authors of study [25] trained Convolution Neural Network (CNN) model over [24]'s dataset to classify a tweet as sexist, racist, and neither. They used Word2vec, character 4-grams, random word vectors, and combinations for training classifiers for their solution. Another interpretable and state of the art multi-view SVM approach is used in [12] to classify hate or no hate over four different datasets (HatebaseTwitter: [21], Stormfront: [16], TRAC Facebook: [15], HatEval: [18]). Word level uni to 5-grams and Character level uni to 5-grams TF-IDF features were used for the experiments.

Regarding the case of hate speech detection, it is quite evident that the majority of classifiers performance is affected due to the inappropriate and low-quality dataset. Therefore it has been addressed in the study and all such issues discussed in section I are considered and resolved through dataset construction. Despite all dataset related issues there are some important concerns which are not considered when text mining, NLP and IR related approaches are used for hate speech detection. Therefore produce large number of false positives. The input text usually have some long-distance relationships, which may be occurred in non-consecutive words. They could not be captured through commonly used features e.g.: n-grams, m-skip- n-grams, etc. Therefore such important syntactic information like, Subject-Object relations or more general Governor-Dependent relations which could easily be captured through dependency tuples, such as; 1. "these black american women are lower class pigs" gives nsubj(women, pigs), 2. "jews by any means, are bull shits in this world" gives nsubj(jews, shits). These basic dependencies could be used for extracting important dependency tuples. There is another issue related to text mining based features which all produce very high dimensions. Dimensions of word n-grams, m-skip-n-gram, character n-grams, and dependency tuples bi-grams, etc. are extremely high. Dimensionality reduction algorithm, i.e.: latent semantic analysis (LSA) which converts high dimensionality information to a "semantic" space of low dimensionality by identifying synonymy and polysemy. In addition to dimensionality reduction, it classifies the information semantically which increased classifier performance. Similarly incorporating appropriate features reduce classification issues. Examples of such features are Extended Named Entity Recognition, Dependency Tuples, etc. These all shortcomings are therefore considered in our solution for better performance.

III. DATASET CONSTRUCTION

The majority of challenges we discussed earlier in the introduction section I belong to the availability of high quality, standard dataset. The dataset should be balanced and multi-classed, tagged by experts under specified definition and clear rules, with the strong agreement of annotator's, furnished with language subtleties. These all are targeted in this section. High performing and effective hate speech detection system highly depend on standard and appropriate dataset. Therefore it is expected to be accomplished in this study.

Initially, five twitter-based popular datasets which are developed for different types of hates are taken and all are re-labeled by experts using 10 classes (see table 3) un-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101977, IEEE Access

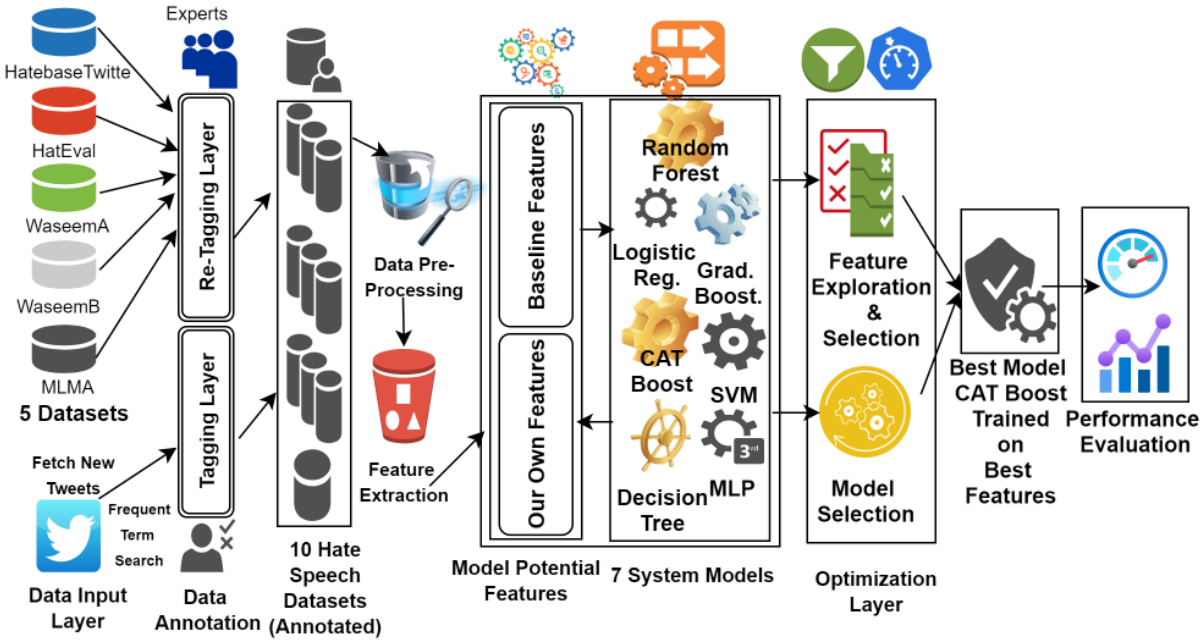

FIGURE 1. High-level System Framework.

der comprehensive definition and detail rules defined in [9]. Tweets outside the scope of 10 classes are tagged as other hate. These five datasets are 1. HatebaseTwitter [21] Tweets:24802, Classes: Hate/Offensive, 2. HatEval [18] Tweets: 10000, Classes: Women/Immigrants, Aggressive/Not, Individual/Group, 3. WassemA [24] Tweets: 16914, Classes: Sexist/Racist 4. WaseemB [34] Tweets: 4033 Classes: Sexist/Racist, 5. MLMA [37] Tweets: 5647, Classes: Gender/Sexist/Religion/Disability. Accumulation of these datasets resulted in a total of 61396 tweets. Frequent terms are then extracted from these re-labeled tweets where each one of them was thoroughly examined and confirmed by the experts for all 10 hate categories. The next phase is started for further hate speech-related tweets collection from Twitter. Tweets were searched containing top extracted hate speech terms. To scrap a large number of tweets in a short span of time, without hitting Twitter API restrictions, a python library "GetOldTweets3" is used. In addition to 61396 tweets from five datasets, 60000 more tweets from Twitter, containing these terms of hate tendencies on 10 hate categories are also collected (see table 1). All these newly fetched tweets are also labeled by the experts, similarly. Dealing with the degree of complexity associated with the problem, and to enhance the performance of classifiers, separate binary classified datasets for each hate category are developed. This method of dataset construction is named 'Binary Classified Multi-Category Hate Datasets'. Each dataset with respect to its category has a clear distinction between what is hate speech and what is not, which provides a fine-tuned line between the both. The dataset is fully enriched with language subtleties. It includes such aggressive, offensive and abusive examples which are clear from hate speech. Similarly those examples are also included which are categorized as hate speech but clear from profane, abusive and offensive language. It simply enables the classifier to reduce misclassification and increase performance.

A. DATA ANNOTATION

To develop the ground truth (GT), 12 experts were asked to manually annotate the data of 121,396 tweets in 6 months. They were told to annotate data into 10 hate speech categories, which can be seen in table 3, together with (Yes, NO, In Doubt) options. The annotators were given detailed guidelines and specific definition as specified in [9], to label the tweets. Since each tweet is annotated by 4 persons and there were a total of 3 teams, we used majority voting to identify the final label. Ground truth (GT) was developed using three unique categories. These distinctive ground certainties were: GTavg, GT4YES, GT3YES

GTavg: In this strategy, every YES answer is allocated two, In-Doubt is allocated one, while NO answer is allocated zero points. Four assessment scores are gathered and a complete score is determined. A tweet may have a score somewhere in the range of zero to eight. If the score is more than four, it is characterized as YES, otherwise NO.

GT4YES: If a tweet is addressed YES by everyone then that tweet is characterized as YES else, it is marked as NO.

GT3YES: If a tweet is addressed YES by three of the experts, it is counted as YES for this model else, it is categorized as NO.

Table 2 gives both, agreement/overlap (which is represented as O) values between GTs and expert answers. Similarly,

TABLE 1. 1. In first phase of our datasets construction, following five twitter-based popular datasets were used. They were developed for different types of hates. They all were re-labeled by our experts using 10 hate classes. Tweets outside the scope of 10 classes were tagged as 'other hate'. Accumulation of these datasets resulted in a total of 61396 labeled tweets. 2. Frequent terms were extracted from these re-labeled tweets. 3. Tweets were searched from Twitter containing top extracted hate speech terms. In addition to 61396 tweets from five datasets, 60000 more tweets from Twitter, containing same 10 hate categories were also collected. 4. Complete set of annotated tweets compiled which were 1,21,396 in total. 5 Only 19% tweets found hate speech in them which were related to 10 hate speech categories. Therefore in each category, the same %age of corresponding no-hate tweets were also selected, to construct a fully balanced dataset. The final dataset contains 45688 labeled tweets under 10 hate speech categories.

S.No	Datasets	Label Categories	No. of Tweets
1	HatebaseTwitter [21]	Hate/Offensive/Neither	24802
2	HatEval [18]	Women/Immigrants	10000
3	WaseemA [24]	Sexist/Racist	16914
4	WaseemB [34]	Sexist/Racist	4033
5	MLMA [37]	Gender/Sexist/Religion/Disability	5647
Total Tweets Used from Above 5 Popular Datasets			61396
Newly Fetched Tweets from Twitter Using Frequent Terms			60000
Total Tweets Labeled by Experts			121396
Only (19x2)% Tweets Considered for Final 10 Categories Datasets			**45688**

TABLE 2. Agreement (O) and Kappa (K) values between experts and ground truths.

Overlap(O) & Kappa(K)	GTavg(O)	GTavg(K)	GT4YES(O)	GT4YES(K)	GT3YES(O)	GT3YES(K)
African Hate	0.91	0.77	0.85	0.57	0.91	0.75
Arab Hate	0.95	0.73	0.84	0.55	0.94	0.71
Asian Hate	0.89	0.67	0.85	0.50	0.87	0.65
Christian Hate	0.85	0.57	0.74	0.49	0.83	0.58
Islam Hate	0.94	0.59	0.76	0.55	0.93	0.59
Jews Hate	0.77	0.45	0.75	0.41	0.76	0.45
Race Hate	0.75	0.68	0.45	0.64	0.73	0.68
Xenophobia	0.92	0.71	0.88	0.70	0.93	0.69
Gender Hate	0.93	0.69	0.88	0.67	0.94	0.68
Sexual Hate	0.91	0.70	0.86	0.68	0.93	0.69

Cohen's Kappa (which is represented as K) scores showing agreement between expert answers and the GTs. Kappa provides a statistical measure by assuming that the overlap is occurring by some coincidence. It can be seen that for overlap GTavg has the maximum agreement values with expert answers for almost every hate speech category, except for Xenophobia and Gender hate which are also very close to GTavg. Similarly, in the case of Kappa, the best outcomes are gotten with GTavg again, except for Christian hate where GT3YES is a bit higher. Since almost all the results of overlap and Kappa are found best in GTavg therefore it will be selected for our further experiments. At the end of the rigorous data annotation process where 61396 tweets were re-labeled and 60,000 new tweets were fresh labeled. The complete set of annotated tweets which were 1,21,396 in total (either scraped from Twitter or found in five datasets), only 19% tweets found hate speech in them which were specifically related to 10 hate speech categories specified in table 3. Therefore in each category, the same %age of corresponding no-hate tweets were also selected, to construct a fully balanced dataset. The final dataset contains 45688 labeled tweets under 10 hate speech categories (see figure:2).

IV. EXPERIMENTAL SETUP

There are a total of ten separate datasets compiled with binary labels each. Different features together with a different set of models are explored over each dataset. Best features are identified and ten independent models are trained. Each tweet will be passed to all ten models and therefore it may have multiple hate classes identified by each model.

A. DATA PRE-PROCESSING

In these kinds of applications, minimal pre-processings are applied, therefore only case-folding and tokenization are simply applied.

B. MODEL FEATURES

It is very important to identify the right approach for you problem first. Therefore it is explored through research studies that using Text Mining, Information Retrieval, or Natural

TABLE 3. Hate categories for Datasets construction.

Hate Labels (Category)	Description of targets
African Hate (Ethnic)	This hate is related to the Africans Americans.
Arab Hate (Ethnic)	Fear or scorn of, or promotion of genocide of Arab individuals.
Asian Hate (Ethnic)	This hate mostly includes Chinese, Paki, Indians and Koreans.
Christian Hate (Religion)	This hate involves people who believe in Christianity.
Islam Hate (Religion)	This is mostly of Islamophobia and hates against Muslims
Jews Hate (Religion)	A large majority consists of Antisemitism who have prejudice, or discrimination against Jews.
Race Hate (Racism)	Hate against a particular social group or community, black, white people
Xenophobia (Refugees)	In this hate, people from other countries are disliked.
Gender Hate (Gender)	Discrimination against different genders including male,female etc.
Sexual Hate (Sexism)	Hate against sexual orientation, Homo-sexual, Hetero-sexual etc.

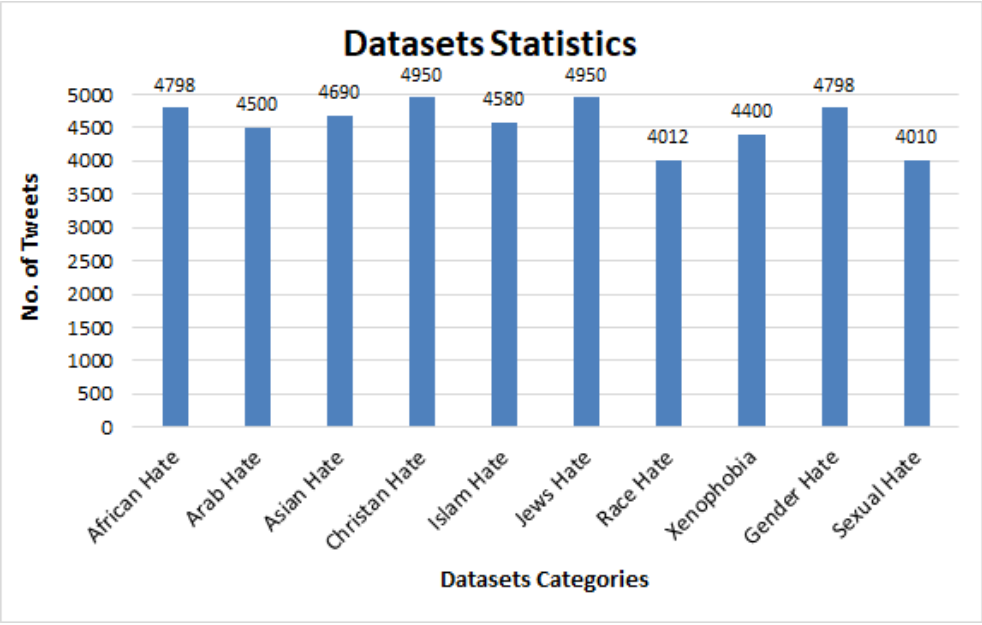

FIGURE 2. Statistics of Ten Datasets.

Language Processing for hate speech identification is considered much effective [9], [14] as compared to Keyword-based, Rule-based/Association Rule Mining [19], Source metadata/Social Network Analysis based approaches [12], [23]. In next phase we examined the most effective and commonly used features reported in text mining related research studies (see section II for these studies). These features are used as baseline for exploration. In addition to these features our own potential features (e.g.: Dependency Tuples, Extended Named Entity Recognition, Features' Dimensionality Reduction, etc.) are also proposed for detail exploration. Finally all these set of features are explored through a verity of combinations and found a few of them most important and efficient in our problem. Following are the features examined for the exploratory study.

1) **Character n-grams(c):**
Different character n-grams, from 2 to 8, are used and found that character 4-grams were most important for the problem. Character n-grams are very efficient for spelling variations which are most common in social media applications.

2) **Word n-grams(w):**
Multiple ranges of word n-grams are explored like 1 to 6 and found word 5-grams as most productive for our experiments. Long n-grams are extremely important for capturing hate phrases and all related words associated with key hate terms, e.g.: "bloody dirty american women shit", "send them back their home", "They must be hung", "kicked them all out", etc. Word n-grams are found much better than the simple bag-of-word (BOW) model, which could only identify key terms

which increase false positives. Character n-grams and Word n-grams both are weighted using TF-IDF vectors [22].

3) **Sentiment Score(Sen):**

VADER, which is specially optimized for social media text, is used for sentiment score identification. It is observed that negative sentiments are common in majority of hate speech categories.

4) **Extended Named Entity Recognition(NER):**

CoreNLP's [17] NER is used, which could recognize (using TokensRegexNERAnnotator sub-annotator): Religion, City, State/Province, Country, Nationality, Job Title, Ideology, Criminal Charges, etc. in addition to other normal Named Entities (Person, Location, Date/Time, Organization, etc.).

5) **General Statistics(GS):**

following general statistics are also used as features for hate speech classification: **Ratio of Capital Letters**, **Text Length**, **No of Words**.

6) **POS Tags(POS):**

CoreNLP's [17] Stanford POS Tagger using GATE module plugin, which is specialized for twitter's data, is used for POS Tags generation, and are used as one of the model features. It is explored that most POS Tags related to hate were: Verb(VBN), Adverb (RB), Adjective (JJ), and Noun (NN/NNS), etc.

7) **Dependency Tuples(Dep):**

Important syntactic information like, Subject-Object relations or more general Governor-Dependent relations which may have long-distance relationships or may be occurred in non-consecutive words, which are not captured through n-grams, could easily be captured through these dependency tuples, such as; 1. "these black american women are lower class pigs" gives nsubj(women, pigs), 2. "jews by any means, are bull shits in this world" gives nsubj(jews, shits). CoreNLP's [17] Basic Dependencies are used for extracting important dependency tuples, and their extracted terms are used as normal bi-grams.

8) **Count of 1st and 2nd Person Pronouns(Pro):**

These two features are also used, which include: occurrence of (I, me, my) and (you, your) and (we, us, our). The presence of these pronouns without NER tags, for example, Religion, Nationality, etc., and certain terms: Refugees, Gays, Women, etc. means no hate speech found.

9) **Dimensionality Reduction:**

Dimensions of word n-grams, character n-grams, and dependency tuples bi-grams, are extremely high. They all are in the form of TF-IDF vectors, developed by scikit-learn's TF-IDF Vectorizer. There is a popular variant of singular value decomposition (SVD) called Truncated SVD, which is applied for dimensionality reduction. When it is applied to term-document matrices which are developed by scikit-learn's TF-IDF Vectorizer, then it is known as latent semantic analysis (LSA). It converts such sparse matrices of high dimensionality to a "semantic" space of low dimensionality by identifying synonymy and polysemy. There is a parameter of Truncated SVD, known as n_components which was set to 500.

C. FEATURE EXPLORATION:

There is an extremely important role of separate and balanced datasets construction for each hate category to solve the challenging problem of hate speech classification. This approach of dataset construction may be called 'Binary Classified Multi-Category Hate Datasets'. Each dataset with respect to its category has a clear distinction between what is hate speech and what is not, which provides a fine-tuned line between the both. All such issues could be explored through t-Distributed Stochastic Neighbor Embedding (t-SNE) plots. It is used for exploring high-dimensional data because it is a non-linear dimensionality reduction algorithm. Using our potential list of features (see section IV-B) different datasets are plotted through tSNE and few important plots are presented in figure 3 for exploration and analysis. Corresponding to the case of intra-dataset separation, our dataset construction approach and potential features are well enough to clearly separate the binary classes with each dataset. It could be explored through figure 3 c that the Gender Hate dataset is clearly separable. Considering the case of inter-dataset separation, many datasets are well separated through potential features, as seen in figure 3 a, and figure 3 d. In both examples, different hate categories (i.e.:Jews vs Gender Hate and Gender vs Sexual Hate) are well separated through the potential list of features. There are few hate groups that are naturally quite overlapping, and therefore they are much difficult for classification as well and may cause an increased level of misclassification, like Race vs African Hate and Race vs Xenophobia (see figure 3 b Race vs Xenophobia). Analyzing these cases following few possibilities are found, for example, appropriate fine-tune features are needed, complex and non-linear models will be required. The complete picture could be seen in figure 4, representing the tSNE plot between all hate categories. Both cases: clearly separable and overlapping are fully distinguished.

D. FEATURES SELECTION:

In the previous section many candidate or potential features were discussed (see section IV-B but few features performed better than others. Optimality of this list of features over all the datasets were explored through different combinations and few combinations were found much effective for hate speech detection. Each group of features was evaluated for F1 and AUC scores, as they are the most suitable and much considered in such problems. Results of some important features exploration using some combinations could be seen in table 4, though many other possible combinations are also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101977, IEEE Access

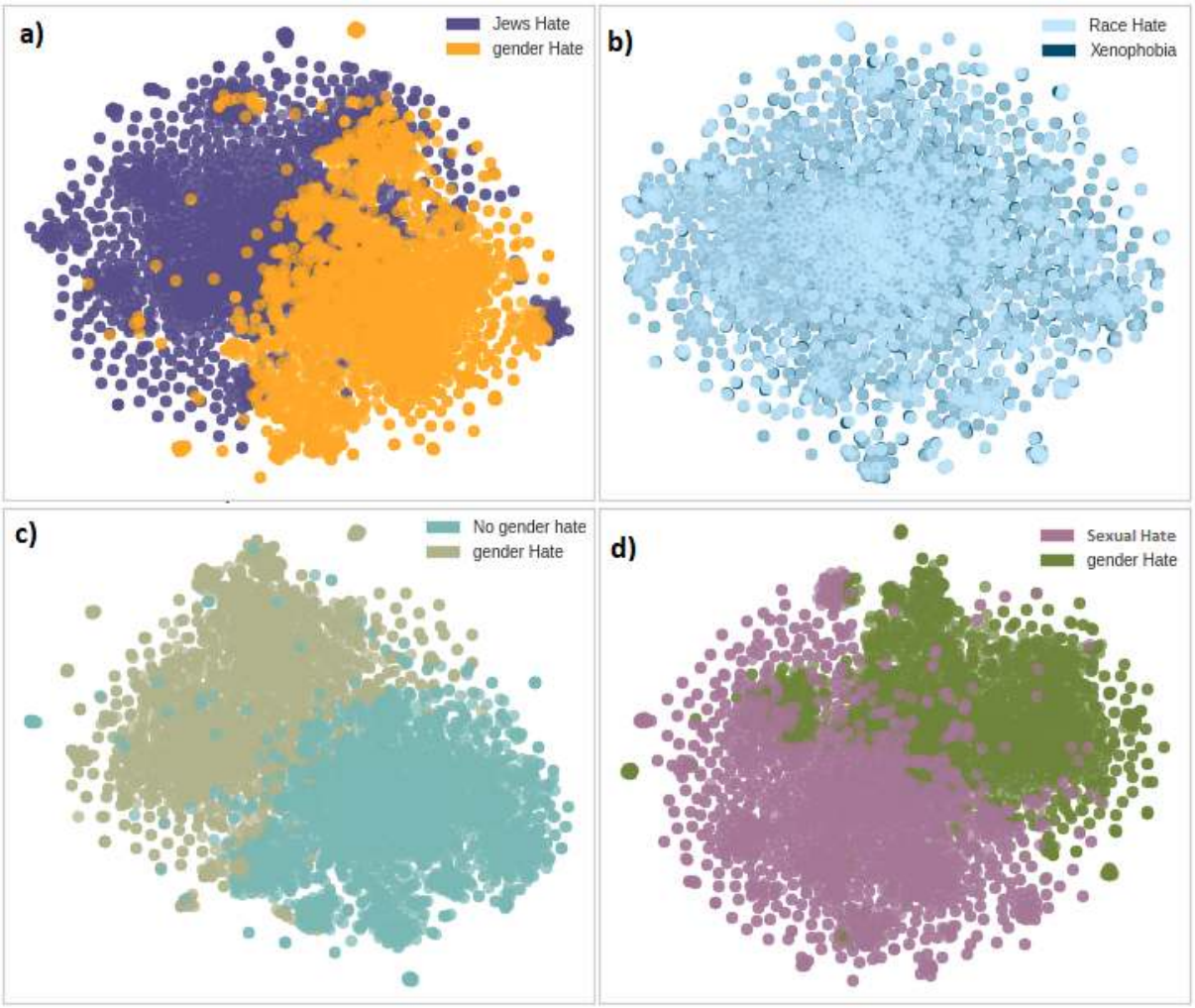

FIGURE 3. These tSNE plots 3 **a,b,c,d** are generated over different datasets, using potential features for exploration and analysis. Different aspects of **1.Intra** vs **2.Inter dataset separation** and **3.overlaping** dataset cases, are explored:
In figure 3 **c**, **Gender Hate dataset** is clearly separable. **1.)** It is the case of **intra-dataset separation**, where the dataset construction approach and potential features are good enough to clearly separate the binary classes.
2.) In the case of **inter-dataset separation**, many datasets are well separated through potential features, as seen in figure 3 **a**, and figure 3 **d**. In both examples, different hate categories (i.e.: in 3 a **Jews vs Gender Hate** and 3 d **Gender vs Sexual Hate**) are well separated through potential list of features.
3.) Some hate groups are naturally quiet **overlapping**, they are much difficult for classification with increased level of mis-classification, e.g.:figure 3 **b, Race vs Xenophobia** or **Race vs African Hate** (this specific case is **not shown** in figure).

explored. The baseline is shown first (see the first column of the table), which is comprised of two important features only: character 2 to 4-grams and word 1 to 5-grams (in short, represented as, n-grams (c+W) in the table). Next, all features are experimented (see the last column of the table) with different machine learning models and found that outcome was increased only in fractions when compared to baseline. In the next stage Dependency Tuples (Dep) are also added to the baseline and significant improvement is seen. It is called the second baseline of the experiment. Adding Part-of-speech(POS) tags to the second baseline, reduced the performance. It means that POS tags are not a good contributor to hate speech detection, therefore it is omitted from further exploration. For the next stage Named Entity Recognition(NER) and Sentiment Score is added in the second baseline, but it did not work and results were slightly reduced, though they were a bit better than previous. Finally, the NER feature is replaced with the compound feature, named: count of 1st and 2nd Person Pronouns(Pro). This group of features produced the best results (see results in boldface, second last column) within all sets of experiments, and the best set of features for hate speech detection are identified. It has already been analyzed that fine-tune features were specially required for overlapping and complex hate classification cases (see section IV-C). Therefore it is explored in this section, that except few features, e.g.: NER, POS Tags, and General Statistics(GS) related features, all other features are found much contributing and therefore presented as proposed features of hate speech detection. These proposed features are; 1. Character 2 to 4-grams, 2. Word 1

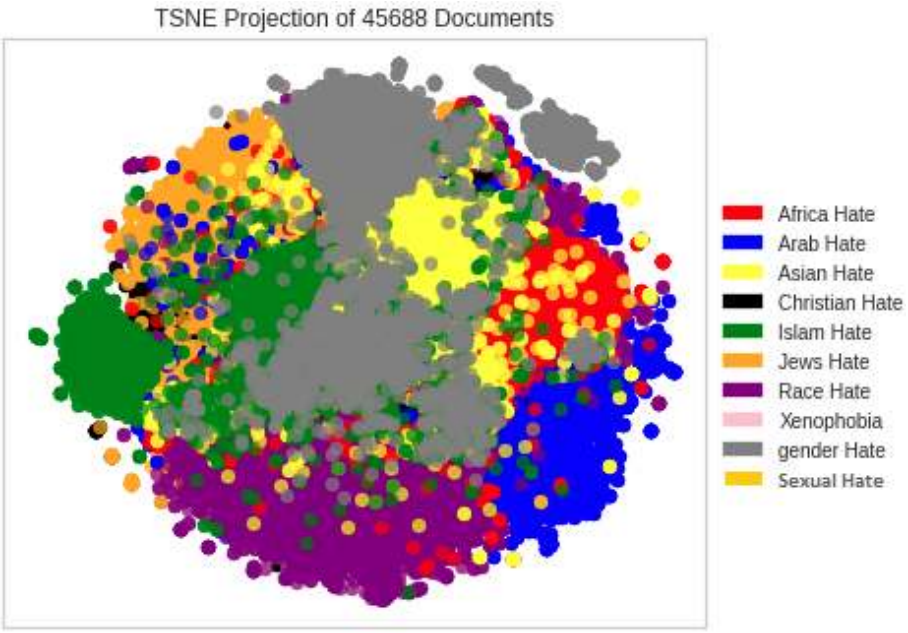

FIGURE 4. tSNE plot of all 10 datasets using potential features, for exploration and analysis. The complete picture representing tSNE plot showing both cases: clearly separable and overlapping, these cases are fully distinguished in the figure.

TABLE 4. F1 and AUC scores generated by CAT Boost using different feature combinations including baseline features. The best scores were generated by Word 1-5 grams, Character 1-4 grams, with Dependency Tuples, Sentiment Scores, and 1st, 2nd Person Pronoun Features. The scores are shown in the boldface under respective F1 and AUC columns.

Datasets	n-grams (Ch+Wrd)		n-grams +Dep		n-grams +Dep +POS		n-grams +Dep +NER +Sen		n-grams +Dep +Sen +Pro		All	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
African Hate	84.1	85.01	86.5	86.9	85.3	86.03	85.8	86.73	**88.2**	**89.9**	84.03	85.43
Arab Hate	84.3	85.22	86.7	87.8	85.02	86.02	85.74	86.9	**88.0**	**89.1**	84.42	85.42
Asian Hate	83.6	84.04	85.9	87.01	84.09	85.9	84.09	86.01	**87.9**	**88.8**	83.99	84.97
Christan Hate	84.7	85.04	86.4	86.8	85.01	85.7	85.71	85.9	**88.3**	**89.5**	84.71	85.3
Islam Hate	85.2	86.21	87.7	88.01	86.3	86.9	86.43	87.01	**89.4**	**90.1**	85.33	85.78
Jews Hate	83.8	84.09	84.9	85.09	83.6	84.6	83.9	84.9	**87.8**	**89.1**	82.95	83.86
Race Hate	81.3	82.01	83.6	85.03	82.4	84.09	82.7	84.8	**85.1**	**86.4**	81.9	82.97
Xenophobia	82.8	83.02	84.1	85.7	83.04	83.9	84.01	84.7	**86.8**	**87.6**	82.94	83.04
Gender Hate	82.6	84.07	85.6	86.07	83.8	85.04	83.9	85.8	**87.7**	**88.6**	82.69	84.67
Sexual Hate	84.5	85.01	86.08	86.92	84.6	85.97	83.7	85.9	**88.2**	**89.7**	84.47	85.8

to 5-grams, 3. Dependency Tuples, 4. Sentiment Scores, and 5. Count of 1st and 2nd Person Pronouns.

E. MODEL SELECTION:

To propose the best solution, different features are explored in the previous section IV-D and a set of popular machine learning algorithms are also evaluated on all datasets. The same set of machine learning algorithms over all datasets are used for the proposed best features. This set of machine learning models used in all experiments will be discussed in this section.

In the problem of hate speech classification, algorithms from different classes are applied, which include: linear, non-linear, tree-based, non-tree based, non-parametric, large margin classifiers, and Ensemble (boosting, election mechanism) models. These algorithms include Support Vector Machine (SVM), Logistic Regression, Multi-Layer Perceptron (MLP), Random Forest, Gradient Boosting Classifier, Decision Tree, and CAT Boost. Following are the briefings of these ML algorithms:

TABLE 5. Using selected best features (n-grams(c+w)+Dep+Sen+Pro), the Accuracy, F1, and AUC average scores at all datasets, with models comparison and their final set of optimal parameters.

Models	Parameters	Avg. Values of Datasets		
		Acc	F1	AUC
Logistic Regression	C=0.01, penalty='L2' , solver='lbfgs', multiclass='multinomial'	81.73	80.49	84.05
MLP	hiddenlayersizes=(500, 1000, 500), maxiter=1000	79.06	77.9	81.2
Decision Tree	max_depth = 20, min_sample_split= 5, ccp= 0.0	82.79	81.59	85.01
SVM	LinearSVC (C=0.001)	81.02	80.41	84.03
Random Forest	max_depth=15 , measure=gini, random_state =0, min_split=5, estimator=500	86.45	85.53	86.76
Grad. Boosting	lr =0.1 , max_depth=15, random_state=0 , n_estimators=500,	88.78	86.04	87.69
CAT Boost	random_seed=50, border_count=110, l2_leaf_reg=7, iterations=1000, learning_rate=0.8, depth=15.	**89.03**	**87.74**	**88.88**

1) Logistic Regression:

A popular machine learning algorithm classified as probabilistic and linear model, used in classification and it uses the categorical class variable. It has different variants, depending on the class variable; binomial, multinomial, or ordinal. It finds the best fit model that can clearly describe the relationship between the dependent and independent variables (see equation 1).

$$P(Y_i = 1|X_i) = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} \tag{1}$$

2) Gradient Boosting Classifier:

It is ensemble learning method of boosting type (e.g.: CAT Boost, Gradient Boosting, etc.). Classified as non-linear, tree based model. Computationally less expensive than ensembles-bagging models e.g.: Random Forest. It normally builds the model in repeated series of rounds. It maintains a set of weights for the training sets. In the beginning, all weights are set equal. Upon each iteration, weights of incorrectly classified examples get increased, so it ultimately gets focused on more hard examples available in the training sets (see equation 2).

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t\left(x\right)\right) \tag{2}$$

3) Decision Tree- Classification and Regression Tree (CART):

CART is used to implement the decision trees which are classified as non-linear, and tree based model. CART is the combination of Classification and Regression Trees. Classification trees work on categorical class variables while Regression trees work on the continuous class variables. In this way, CART has the ability to predict the class variable whether it is categorical or continuous [40].

4) Random Forest

It is ensemble learning method of bagging type. It is simply called voting system and classified as non-linear, tree based model. Random Forest has been used in this study for classification purposes because it has the ability to build the model based on the combination of tree predictors where each tree depends on the value of random vector sampled independently and it follows the same distribution for all the available trees in the forest (see equation 3 which describes the marginal function of random forest).

$$mr(X,Y) = E_\Theta[I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{j}(X,Y)] \tag{3}$$

5) Support Vector Machine Classifier

SVM are known as large margin classifier, and classified as liner and non-linear both. In support vector classification, the model is built by nonlinear mapping of input vectors on high dimensional feature space, which in turn is used to construct the linear decision surface. This decision surface has key importance towards the generalization ability of machine learning [41].

All these sets of models are evaluated over all datasets separately and then the single average score is computed for each model to make comparison easy. Their results could be seen in table: 5 with all hyperparameter configurations. Results are evaluated over popular measures, in terms of Accuracy, F1, and AUC scores. Python using scikit-learn is used as our programmable language throughout the experiments.

After applying the dimensionality reduction, this problem is benefited from the utilization of many state-of-the-art models. It has already been analyzed in section IV-C that some complex cases of hate classification could correctly be identified through complex non-linear models. Therefore it could be observed that specific group of advanced non-linear classifiers, such as, Decision Tree, Random Forest,

Gradient Boosting, and CAT Boost performed much better, respectively than non-linear models, like Support Vector Machine (SVM), and Logistic Regression, whereas Multi-layer Perceptron (MLP) is an exception which performed worst among all. The best performing model was CAT Boost, whose results against all individual datasets are shown in the table: 4.

V. RESULTS AND DISCUSSION:

The need for appropriate fine-tuned features and complex non-linear models has already been explored and recommended in section IV-C. Therefore appropriate proposed features have already been identified (see section IV-D) and utilized in appropriate models (see section IV-E). Considering the overlapping or complex cases of hate speech classification, by exploiting the proposed features, the best model has provided much better results. Which is in full compliance with the findings of section IV-C. It concludes that using non-linear model with appropriate fine-tuned features could produce much better results. This could be seen through individual results of jews, race, and xenophobic hate datasets in table 4). These results are now much improved and also very close to other category datasets. Considering the AUC measure, results produced by baseline model were 84.0, 82.0, 83.0 respectively for jews, race, and xenophobic hate datasets, while results produced by best feature model are 89.1, 86.4, 87.6 respectively, after improvement.

It could also be observed through our explorational study, that among complete list of potential features which simply combines tow basic group of features. First group includes baseline features which are commonly used and effective features (e.g.:Character n-gram, Word n-gram, General Statistics, Sentiment, etc.) recommended in related studies. Second group includes our own presented new features (Dependency tuples, Extended NER, etc.). It is explored that only few features (n-grams, Sentiment, Pronouns) from first group and just one feature (Dependency Tuples) from second group are found best performing. It is concluded after deep analysis of table 4 that except few features, e.g.: Extended NER, POS Tags, and General Statistics(GS) related features, all other features are found much contributing and therefore proposed for final model.

This is quite evident, that even all top three models are non-linear, tree-based, and state-of-the-art ensemble learning algorithms (see table:5). In terms of performance, these models are evaluated under three suitable measures. F1 is more useful than Accuracy, and more suitable when you have an imbalance class distribution. It is basically the weighted average of Precision and Recall. Unlike Precision and Recall which are class-based, Accuracy is overall system-based. It is good when we have a balanced class distribution. Since our datasets are slightly imbalanced, therefore AUC is also used, to validate the performance of machine learning algorithms. AUC is more statistically consistent and more discriminating than F1 and Accuracy.

The most popular and advanced machine learning model, CAT Boost has shown the best average scores, in terms of Accuracy, F1, and AUC, which were 89.03, 87.74, and 88.88, respectively (see table 5). These results seem quite appreciating, considering the context of the hate speech problem's criticality. Similarly, the Gradient Boosting model performed next to CAT Boost with minor difference, which scored 88.78, 86.04, 87.69 under the same measures of Accuracy, F1, and AUC, respectively. Random Forest stood at the top 3rd with slight variation in scores, which are 86.45, 85.53, and 86.76 corresponding to Accuracy, F1, and AUC, respectively (see table:5). It could also be seen in table 5 that Multi-Layer Perceptron (MLP) performed worst as compared to other machine learning models, whose scores were: Accuracy=79.06, F1=77.9, and AUC=81.2. Both linear models, Support Vector Machine (SVM) and Logistic Regression, gave an average performance. Their scores are almost similar. Logistic Regression scores were: Accuracy=81.73, F1=80.49, and AUC=84.05. In terms of Accuracy, F1 and AUC, the SVM produced scores as: 81.02, 80.41, and 84.03 respectively. It is also observed during the experiments that these linear models reported poor results for overlapping hate categories, such as Race and African hate datasets. It confirms that linear model are unable to develop complex decision boundaries even in the presence of appropriate features.

Finally combined dataset approach has also been explored. In this approach single dataset was constructed by combining each dataset with its true and false labels. The single classifier was trained over a combined dataset using one vs all. CAT Boost was still best performing. It gave an AUC score of 88.8, as shown in figure 6 (see word cloud picture of the combined dataset in figure 5).

There were some misclassification cases seen which were classified as no-hate speech but they actually belong to some hate category, e.g.: "what you need just a lipstick and a wig and be who are". These cases are mostly an example of Context-Aware Hate Speeches because as an individual sentence they could never be considered as hate speech, but considering the complete context, like it was commented for young boys and their gender identity is targeted, only then they are identified as hate speech. There are some other examples in which different analogies are used therefore they become challenging for hate speech identification, e.g.: "Frog is calling", here Muslims are accused of their prayer's call.

A. CRITICAL ANALYSIS:

Performance of the hate speech detection system, presented in this study, is compared with two related research studies, also presented in section II. These studies are following:

In first study [26] character n-grams (2-8), word n-grams (1-3), and word skip-grams (1-,2-,3- bi-grams) were used as features over a multi-class SVM model. It shows that the use of character 4-grams helps in accomplishing the best 78.7% accuracy. In the second study, [12] state of the art multi-view SVM approach was used. It also provides interpretable outcomes. Word level uni to 5-grams and Character level uni

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3101977, IEEE Access

to 5-grams TF-IDF features were used for the experiments. The model accomplished 80.3% accuracy. The comparison of these studies is shown in table 6. Our model achieved much better results when compared with both studies. It scored 89.0 in terms of accuracy.

In section IV-D different baseline features have already been explored to identify fruitful features for multi-class hate speech detection. These features were found most effective and commonly used among many similar studies. The basic group of baseline features is also considered for comparison. It could be analyzed that CAT Boost has provided much-improved results over multi-class hate speech datasets as compare to related studies. It shows an accuracy score of 85.3, which is quite better than 78.7 and 80.3 which are produced by studies [26] and [12], respectively. Though the result produced by our final proposed features modal is much higher than all these studies, which has achieved 89.0% accuracy.

FIGURE 5. Word cloud of hate speech combined dataset.

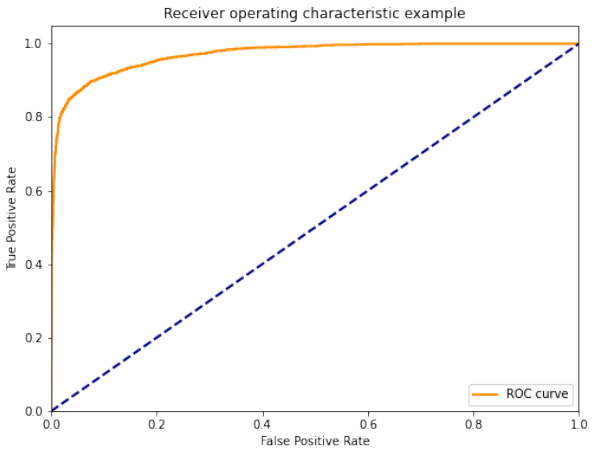

FIGURE 6. Combined dataset result for best performing CAT Boost model which gave AUC score of 88.8 which is represented in the graph.

VI. CONCLUSION

In this study, major challenges are identified first and the complex problem of multi-class automated hate speech classification for text is accomplished with much better results. Ten separate binary classified datasets consisting of different hate speech categories are constructed. Each dataset was annotated by experts with the strong agreement of annotators under comprehensive, clear definition and well-defined rules. Datasets were well balanced and broad. They were also supplemented with language subtleties. Compilation of such dataset was achieved as necessary requirement for filling the gap of the field. After the development of high-quality datasets, a list of effective, commonly used and recommended features extracted from related studies under the field of text mining were identified. In addition to these features our own potential features were also proposed. These features were then explored and identified with respect to their problem objective. It is found that character 2 to 4-grams, word 1 to 5-grams, dependency tuples, sentiment scores, and count of 1st, and 2nd person pronouns were very effective. Latent Semantic Analysis (LSA) as a dimensionality reduction algorithm was also applied and found much effective in such high dimensional classification problems. Datasets were completely explored through tSNE multi-dimensional plots. These plots identified issues like the need for appropriate discriminating features, complex data overlaps, and non-linearity. Therefore complex, non-linear models were used for classification, and the most popular and advanced machine learning model CAT Boost was found top-performing over all datasets. CAT Boost has shown the best average scores, in terms of Accuracy, F1, and AUC, which were 89.03, 87.74, and 88.88, respectively. These results seem quite appreciating, considering the context of the hate speech problem's criticality. Similarly, the Gradient Boosting model performed next to CAT Boost with minor difference, which scored 88.78, 86.04, 87.69 under the same measures of Accuracy, F1, and AUC, respectively. Random Forest stood at the top 3rd with slight variation in scores, which are 86.45, 85.53, and 86.76 corresponding to Accuracy, F1, and AUC, respectively (see table:5). The performance of the final model is also compared with two related studies and our initial baseline. It is worth mentioning that the model outperformed all these.

TABLE 6. Comparison of our best model CAT Boost, trained on final proposed features, with two related studies and our studys' baseline.

S.No.	Research Study	Best Model	Features	Acc	F1
1	Malmasi et al.([26])	Linear SVM	Ch and Word n-grams, Word skip n-grams	78.7	77.2
2	Sean MacAvaney et al. ([12])	Multi view SVM	Ch and Word n-grams	80.3	80.3
3	**Our Studys' Baseline**	CAT Boost	Ch and Word n-grams	**85.3**	**83.7**
4	**Our Study**	CAT Boost	Ch, Word n-grams, DepTup, Sent, Pro	**89.0**	**87.7**

VII. FUTURE DIRECTION

In the future, it is decided to expand our horizons and promote effective measures to further strengthen our research. It will be done through longitudinal and latitudinal expansion in datasets. For example: Reducing the misclassifications and increasing clarity and better understanding for classifiers, precise examples or cases will be added. This will be done specifically for complex and overlapping hate speech categories. Add other hate speech categories in form of datasets, etc. Regarding models: Appropriate deep learning models (e.g.: BRNN's LSTM and GNU, Transformers, GAN, etc.) for context-aware and multi-modal hate speech detection (e.g.: CNN, etc.), will be explored.

VIII. ACKNOWLEDGEMENT

Having the role of experts: Syed Zohaib Abbas, Abdul Samad, Mehboob Elahi, M. Umar, Aasma ,Areeba, Aun Raza , Muneeb Ahmed, Saad Ahmed Ansari, Owais-ur-Rehman, Nayyar Ahmed, Arbaz Khan (Department of Computer Science, DSU, Karachi), have labeled the datasets.

REFERENCES

[1] 2019 hate crime statistics. Retrieved from: https://ucr.fbi.gov/hate-crime/2019, Accessed 2021
[2] ILGA. 2016. Hate crime and hate speech. Retrieved from http://www.ilga-europe.org/what-we-do/ouradvocacy- work/hate-crime-hate-speech.
[3] Twitter. 2021. The Twitter Rules. Retrieved from https://support.twitter.com/articles/.
[4] Youtube. 2021. Hate speech. Retrieved from https://support.google.com/youtube/answer/2801939?hl=en.
[5] Facebook. 2013. What does Facebook consider to be hate speech? Retrieved from https://www.facebook.com/help/ 135402139904490.
[6] Christian Wigand and Melanie Voin. 2020. Speech by Commissioner Jourová—10 years of the EU Fundamental Rights Agency: A call to action in defence of fundamental rights, democracy and the rule of law. Retrieved from http:// europa.eu/rapid/press-release_SPEECH-17-403_en.htm.
[7] Nockleby JT. Hate Speech. Encyclopedia of the American Constitution. 2000; 3:1277–79.
[8] Wermiel SJ. The Ongoing Challenge to Define Free Speech. Human Rights Magazine. 2018; 43(4):1–4.
[9] Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51.4 (2018): 1-30.
[10] No hate speech movement. 2021. No hate speech movement. Retrieved from https://www.nohatespeechmovement.org/.
[11] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and IngmarWeber. 2016. Analyzing the targets of hate in online social media. arXiv Preprint arXiv:1603.07709 (2016).
[12] MacAvaney, Sean, et al. "Hate speech detection: Challenges and solutions." PloS one 14.8 (2019): e0221152.
[13] Ivana Kottasová. 2017. Europe says Twitter is failing to remove hate speech. Retrieved from http://money.cnn.com/2017/06/01/technology/twitter-facebook-hate-speech-europe/index.html.
[14] Schmidt, Anna, and Michael Wiegand. "A survey on hate speech detection using natural language processing." Proceedings of the fifth international workshop on natural language processing for social media. 2017.
[15] Kumar R, Ojha AK, Malmasi S, Zampieri M. Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). ACL; 2018. p. 1–11.
[16] de Gibert O, Perez N, Garc'ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive Language Online@EMNLP; 2018.
[17] Stanford CoreNLP 4.2.0; Available from: https://stanfordnlp.github.io/CoreNLP/, Accessed 2021
[18] CodaLab—Competition;. Available from: https://competitions.codalab.org/competitions/19935.
[19] Haralambous, Yannis, and Philippe Lenca. "Text classification using association rules, dependency pruning and hyperonymization." arXiv preprint arXiv:1407.7357 (2014).
[20] Natalya Tarasova. 2016. Classification of Hate Tweets and Their Reasons using SVM. Master's thesis. Uppsala Universitet
[21] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.
[22] https://monkeylearn.com/blog/what-is-tf-idf
[23] Unsvåg, Elise Fehn, and Björn Gambäck. "The effects of user features on twitter hate speech detection." Proceedings of the 2nd workshop on abusive language online (ALW2). 2018.
[24] Zeerak Waseem, Dirk Hovy "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter (2016)"
[25] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." Proceedings of the first workshop on abusive language online. 2017.
[26] Malmasi, Shervin, and Marcos Zampieri. "Detecting hate speech in social media." arXiv preprint arXiv:1712.06427 (2017).
[27] Vu, Xuan-Son, et al. "HSD shared task in VLSP campaign 2019: Hate speech detection for social good." arXiv preprint arXiv:2007.06493 (2020).
[28] Alper Gun and Pınar Karagoz. A Hybrid Approach for Credibility Detection in Twitter. In Hybrid Artificial Intelligence Systems, pages 515–526. Springer, 2014.
[29] Mathew, Binny, et al. "Analyzing the hate and counter speech accounts on twitter." arXiv preprint arXiv:1812.02712 (2018).
[30] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6 (2018): 13825-13835.
[31] Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv:1809.08651 (2018).
[32] Al-Hassan, Areej, and Hmood Al-Dossari. "Detection of hate speech in social networks: a survey on multilingual corpus." 6th International Conference on Computer Science and Information Technology. Vol. 10. 2019.
[33] Biere, Shanita, Sandjai Bhulai, and Master Business Analytics. "Hate speech detection using natural language processing techniques." Master Business Analytics Department of Mathematics Faculty of Science (2018).
[34] Waseem, Zeerak. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter." Proceedings of the first workshop on NLP and computational social science. 2016.
[35] Wich, Maximilian, Jan Bauer, and Georg Groh. "Impact of politically biased data on hate speech classification." Proceedings of the Fourth Workshop on Online Abuse and Harms. 2020.

[36] Hatebase Lexicon. 2021. Hatebase. Retrieved from https://www.hatebase.org/.
[37] Ousidhoum, Nedjma, et al. "Multilingual and multi-aspect hate speech analysis." arXiv preprint arXiv:1908.11049 (2019).
[38] Jacobs, J. B., and Potter, K. 2000. Hate crimes: Criminal Law and Identity Politics. Oxford University Press.
[39] Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: The 3rd Workshop on Natural Language Processing for Computer-Mediated Communication@Conference on Natural Language Processing; 2016.
[40] Loh, Wei-Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1, no. 1 (2011): 14-23.
[41] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.

KHUBAIB AHMED QURESHI is currently affiliated to Department of Computer Science, DHA Suffa University as Assistant Professor, having 20 years of comprehensive research and teaching experience, continuing research in the area of Computer science named Data Science, Complex Networks, and Social Computing, etc. He has authored several research articles along with chapters in different books.

DR. MUHAMMAD SABIH is working in the field of Computer and Electrical Engineering. He is currently an Assistant Professor in DHA Suffa University and actively engaged in developing solutions from industrial data utilizing machine learning methods for estimation, modeling, and compensation. He has one US patent and around 10 peer-reviewed papers. His current research interest include Industry 4.0, Data Science, Modeling, and Estimation for real world problems.