2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT) | 978-1-6654-5357-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCT56969.2023.10075749

Detecting Hate Speech in Hindi in Online Social Media

Anushka Sharma
Dept. of Information Technology
Indira Gandhi Delhi Technical University for Women
New Delhi, India
anushka023mtechit21@igdtuw.ac.in

Rishabh Kaushal
Dept.of Information Technology Indira Gandhi Delhi Technical University for Women
New Delhi, India
rishabhkaushal@igdtuw.ac.in

Abstract— —**Because of the rise in online hatred, the research communities of artificial intelligence, particularly natural language processing, have been developing models for identifying online hatred. Recently, code-mixing, or the usage of multiple languages in social media conversations, has made multilingual hatred a significant difficulty for automated detection. The crucial task involved in NLP is identifying inciting hatred in writings on social networking sites. This work has several relevant applications, including analysis of sentiments, cyberbullying in online world, and societal & political conflict studies. Using tweets that have been put online on Twitter, we analyze the issue of hatred detection in multilingual functionality in this paper. The tweets have the text annotations and the speech category (Normal speech or Hate speech) to which these belong. We, therefore, recommend a monitored method for detecting hatred. Additionally, the classification approach is provided, which uses certain characters level, words level, and lexicons-based features for identifying hate speech in the corpus. We obtain results of 96% accuracy in identifying posts across four classifiers. Index Terms—Hate speech, Multilingual, Code-mixing, NLP**

Keywords— Hate speech, Multilingual, Code-mixing, NLP.

I. Introduction

Earlier used as tools to connect with friends and family and share their memorable moments, nowadays, social media platforms are being used to disseminate hate speech and other inflammatory content among their users. The potential of online media sites is to connect with and discuss issues with anybody in the world. Social networking has been a powerful platform for immediately spreading hate speech online. Because of its paralinguistic signs used in social media posts, such as emoticons and hashtags, and a large amount of poorly written content they contain makes it very complex for automatic identification. For the last two decades, social media and Internet usage have soared, transforming the way of communication among people. Along with beneficial results, it has offered a lot of risks and detrimental outcomes. The concerns raised due to the relationship between online hate speech and subsequent violence against minorities and other immigrants have increased the role of government and other businesses in their monitoring. Analysts claim that the hate crimes shifting in the political landscape may amplify the tension through social media, and the violence fueled may range from lynchings to ethnic cleansing. Hate speech may be in the form of behavior, speech, or writing using derogatory terms to criticize or refer to a person or a specific group on one's identity; for example, based on their race, religion, nationality, ethnicity, or any other identifying characteristic.

Nearly everywhere in the world, we have documented incidents of hate crimes. Social media is being used by a large number of individuals around the world to communicate; approximately one-third population of the world is using Facebook alone. As per experts, this number is going to rise in the future. Those who are sexist, racist, and homophobic have formed online communities to reinforce and validate their beliefs to be used for inciting violence. Social media platforms also allow these non-state violent actors to publicize their acts.

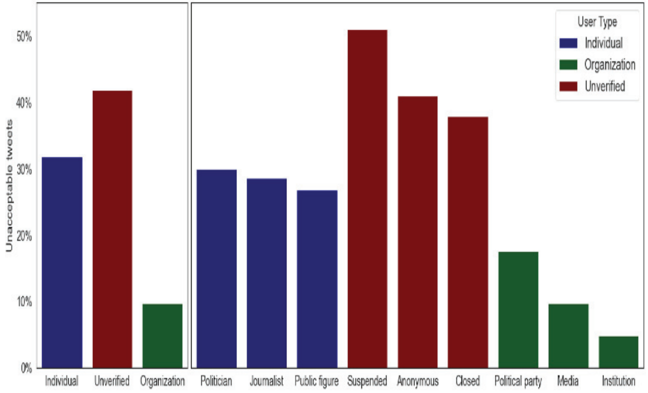


Fig. 1. Communities of retweeters identify the principal source of hateful speech.[9]

[1]A large number of people presently communicate through social networking platforms. As many people have switched online, experts assert that individuals inclined towards misogyny, racism, or homophobia have expressed their views on social media. Research claims that some internet risks, particularly hate speech, still exist for Indian users. Microsoft revealed the results of its 2020 Digital Civility Index and its yearly study, "Civility, Safety, and Interactions Online - 2020." (DCI). Since 2016, The percentage of hate speech has grown by 26%. Many types of modern and anonymous risks are creating their exposure to online users. Approx 50% Indian users have claimed that they have encountered risks through online strangers. India has several official languages, including Hindi. Many activities on social media are carried out using the Hindi language.[2] There have been numerous studies done on

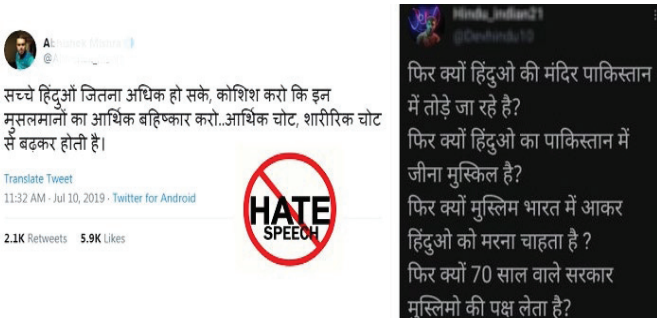
[1] https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[2] https://researchoutreach.org/articles/hate-speech-regulation-social-media-intractable-contemporary-challenge/

978-1-6654-5357-8/23/$31.00 ©2023 IEEE

the issue of investigating hateful speech. In order to identify hate speech, the main focus has remained on developing datasets in a monolingual language like English.

Researchers have also proposed several datasets for multilingual languages, such as Hinglish, to build accurate models to recognize hostile content. Several research works have been conducted to offer new corpus for multiple languages like Hindi, even though most of prior works have been performed on English..



A[redacted]
@[redacted]

सच्चे हिंदुओं जितना अधिक हो सके, कोशिश करो कि इन मुसलमानों का आर्थिक बहिष्कार करो..आर्थिक चोट, शारीरिक चोट से बढ़कर होती है।

Translate Tweet
11:32 AM · Jul 10, 2019 · Twitter for Android

2.1K Retweets 5.9K Likes

[redacted]
@[redacted]

फिर क्यों हिंदुओ की मंदिर पाकिस्तान में तोड़े जा रहे है?
फिर क्यों हिंदुओ का पाकिस्तान में जीना मुस्किल है?
फिर क्यों मुस्लिम भारत में आकर हिंदुओ को मरना चाहता है ?
फिर क्यों 70 साल वाले सरकार मुस्लिमो की पक्ष लेता है?

Fig. 2. Example of Hate Speech against Muslims. [6]

Identifying hate speech is a significant task on online platforms. This uncontrolled spread of hate can extensively hurt marginalized groups, individuals, and society. The dissemination of hate speech through online media is a critical issue. People subjected to this online spread of hate may lose their self-esteem and face anxiety, uncertainty, and fear. The automatic identification of hatespeech can help in dissemination of hateful content on social media.

II. RELATED WORK

For a long time, the research community has been interested in the issues pertaining to hate speech. The underestimation and reduction of respect of minority members are fueled by hate speech that is spoken in public (Greenberg and Pyszczynski, 1985) [2], and repeated subjection to hate speech can heighten a person's bias towards society (Soral et al.,) [23]. To create algorithms to identify hateful content better accurately, Numerous corpora have been proposed by researchers (Waseem and Hovy et al. [25]; Davidson et al., [7]; de Gibert et al., [8]; Kumar et al., [12]). Even though English makes up the vast majority of these datasets Using Transformer-based pre-trained language models, specifically created to construct contextual embeddings of text sequences, Mukherjee et al.(2021) [17] address the problem of Detecting and Removing hate speech.Using four publicly available datasets from various social media sites, they assessed two of these models—RoBERTa and XLNet—and compared them to the industry norms. This analysis demonstrates that the Transformer-based models either outperform or match all of the baseline results obtained by earlier models like the long short-term memory by significant margins (LSTM) and 1-dimensional architectures of convolutional neural network (1D-CNN). Rana et al., [20] concluded that the speaker's emotional state and its impact on the spoken words would be the most crucial characteristic in categorizing hate speech. The first multimodal deep learning system is proposed in this paper to combine auditory data that convey emotion with semantic features to identify hateful content. Our findings show that when detecting hateful audiovisual content, emotional qualities significantly outperform text-based models. As there is currently no dataset for multimodal learning, this research also introduces a brand-new Hate Speech Detection Video Dataset (HSDVD). In order to automatically analyze hate speech in online media, G Kovacs et al. [11] suggested a ´ deep natural language processing (NLP) model integrating recurrent layers and convolutional. They tested their model using the HASOC2019 corpus, and with an F1 score of 0.63, they identified hate speech. However, the DNNs' capacity to learn well also entails a higher risk of overfitting. Especially utilizing the few training data provided (as was the case for HASOC) Malmaison et al. [13] examine the detection of negative comments on social networking platforms while segregating them from other forms of obscenity. They use supervised classification techniques to create lexical baselines for this job using a recently available dataset that has been annotated for it. Their system employs word N-grams, skipgrams, and character N-grams as characteristics. In terms of accurately recognizing posts across three classes, and achieved results of 78Hettiarachchi et al., [10] Their study's objective is to automatically detect hate speech in posts and papers made in romanized Sinhala on social media. Additionally, most research on recognizing hate speech has been carried out in English or the target language; however, this study identifies Sinhala terms that are Romanized Sinhala and written in English characters. In this study, several N-gram values for bigram, trigram, and unigeam were compared along with four machine learning techniques. The Min-Df value was employed. The investigation and comparison of several features for the various classifiers used in the study's classification of hate speech on Facebook. They acquired a dataset of over 2500 comments, some of which contained hate rhetoric, and used a number of attributed to train and test our classifier. Bohra et al. [3] presented a collection of tweets posted online in Hindi and English to examine the issue of hate-speech detection in code-mixed texts. The tweets are written in a language that employs annotations at the word level, and they fall into two categories: normal speech and hate speech. Additionally, they suggest the supervised classification approach to analyze the word-level, character-level, and lexiconbased elements for identifying hatred in the text. Utilizing two publicly released datasets with descriptions for racist, sexist, hateful, or other provocative content on social networking platforms, Mozafari et al. [16] examine the suggested methodology. Compared to other methods, the outcomes show that their approach works superbly on these datasets regarding precision and recall. As a result, the model can account for some biases in the annotation and collection of data and may help us arrive at a more accurate model. Even though English makes up the vast majority of these datasets, new datasets for many languages, including Hindi(Bohra et al., [4]; Modha et al.,) [15], Greek (Pitenis et al.,) [19], Turkish (oltekin) [5] ¨ and Mexican Spanish (Aragon et al., 2019) [24], have already ´ been shared through several recent collaborative tasks. (Mandl et al., 2019 [14] ; Kumar et al., 2020 [12]) Additionally, several models were developed that use these databases. A hold-out test dataset has been used to gauge these models' performance. The research community needs these datasets to build models for recognising hate speech, it is still challenging to identify these models' flaws. As a framework for model evaluation, Ribeiro et al. (2020) [21] developed functional tests in NLP, demonstrating that the method may be used to determine a model's weaknesses and strengths at a detailed parameter that may be frequently hidden by certain high-level measures such as F1- score or accuracy. In order to assess model performance on what Palmer et al. (2020) [18] refer to as "complex abusive language," specifically the employment of adjective nomenclature, reclaimed slurs, linguistic distancing, and adjective nomenclature, three datasets were created. Recently, Rottger et al. [22] modified the said framework ¨ for creating HATECHECK to assess hate speech detection models that cover functionalities of 29 models inspired by discussions with stakeholders of civil society. HATECHECK covers 29 model functionalities. M Das et al. [6] introduce HateCheckHIn, a collection of functional tests, which directly extends earlier work by Rottger et al. (2020). [22]. In this study, we examine how well the fundamental machine learning methods for detecting hate in Hindi text.

III. DATASET DESCRIPTION

Today, it is a standard procedure to employ code-mixing while writing multilingual postings on social media or multiple languages in a single discussion or speech. As a result, when reading a few research articles, we discovered one that researchers at IIT Kharagpur wrote. As a result, A group of functionalities is introduced by Mithun Das et al., [6] for evaluation. They were motivated to create these kinds of functions by actual social media conversations. For each capability, they create test cases using Hindi as a basic language. HateCheckHIn is the name of their adopted dataset, and these datasets were used from Mandl et al. [15], and Bhardwaj et al. [1], which were made available as a result of the shared HASOC-2021 and CONSTRAINT-2021 projects. 6,126 tweets in the Mandl et al. (2021) dataset have been classified as hateful, offensive, profane, or neither. The 8,192 tweets in the Bhardwaj et al. (2020) dataset were divided into categories such as non-hostile, provocative, hate speech, fake news, and defamatory. The data points with one of the two labels "normal" or "hate speech" were utilized for both datasets. These are monolingual language datasets (English).

TABEL I is an example describing the dataset through counts of some key entities involved.

TABLE I. INFORMATION ABOUT THE DATASET.

Label	Counts
Number of Hate Speech	4468
Number of Non-Hate Speech	1416

Every dataset also comprises data attributes. TABEL II describes attributes of data. In the case of supervised learning, 'label gold attribute is the target, and the 'test case' would be considered as the input for predicting the label. The dataset is publicly available at these links.3 4

TABLE II. DETAILS OF DATA ATTRIBUTES

Data Attributes	Brief Explanation
test_case	The text of the test case.
target_ident	The protected group that the test case targets.
label_gold	The test case's evaluation of whether it is hateful or not.

A. Data Pre- processing

The data was pre-processed in the following ways before feature extraction. (1) Merged Datasets: The datasets for the monolingual language (English) and the multilingual language (Hinglish) were combined by removing extra rows and columns and giving the columns in both datasets the same names. (2) Removal of duplicate values. (3) Removal of null values: There were 24 null values in the target ident columns. Therefore, we removed them because our model was unaffected by them. (4) Removal of Stopwords: At a contextual level, stopwords words have nothing to do with the content, and their removal has no impact on the text's overall meaning. (5) Performed stemming and lemmatization: To examine the meaning of a word, we implemented lemmatization and stemming. Lemmatization makes use of the word's context, whereas stemming makes use of the word's stem. (6) POS Tags: Part-of-speech (POS), According to the word's definition and context, words in a text (corpus) are categorized using the common NLP technique known as tagging. (7) Calculate the text's length, punctuation usage, frequency of stopwords, and the quantity of @ symbols in the original (not processed) text for each row. The next four columns now have the headings "length," "punct," "stopwords," and "mentions" were created to record this information.

IV. PROPOSED METHODOLOGY

Machine learning models only comprehend numerical data since they cannot process categorical input. Vectorization is the process of converting categorical data to numerical values. Many vectorization methods exist, including BOW, TF IDF, Glove, and Word2vec. We employed the vectorization techniques BOW and TF-IDF in our baseline experiments. We explain the vectorization techniques in first sub-section, followed by machine learning algorithms in second subsection.

A. Vectorization Techniques

Bag of Words (BoW): BoW is a process of text modeling. The method involves the extraction of the features in a sentence, paragraph, or dataset. It is the simplest method of feature extraction from documents. It is defined as a representation of text describing the occurrence of words in the document. In this, attention is not paid to grammatical conventions or word order; the record is kept only of word counts. Every detail regarding the structure or arrangement of words in a document is neglected. It is not concerned about the appearance of recognized terms in a document; the only interest is whether they do. The text's disorder and lack of structure are critical problems with it. Algorithms for machine learning favour organised and fixed-length inputs. This method is used to convert texts of variable length into fixed-length vectors. Further, machine learning models work with numerical

[3] https://github.com/paul-rottger/hatecheck-data

[4] https://github.com/hate-alert/HateCheckHIn

data compared to textual data. By this method, a sentence gets converted into a vector of integers

Term Frequency and Inverse Document Frequency (TF-IDF): In TF-IDF, the term frequency (TF(t, d)) refers to the repetition with which a term appears in a text. In other terms, ratio of number of times term t appears in document d divided by total terms in the document, denoted by Td.

$$TF(t,d) = (N(t,d)|T(d))$$

Term Frequency alone is insufficient to offer effective solutions. Additionally, we must pair it with another expression known as IDF. The division of a total of sentences in the corpus and sentences containing the word yields to percentage, which is then logarithmically transformed.

$$\mathrm{IDF(t)} \ = \ \log(\{\mathrm{D}(total)|\mathrm{T}(d)\})$$

is obtained by dividing total documents containing t th term, denoted by Dt, with total documents, denoted by Dtotal, and taking log of the output.

B. Machine Learning Algorithms

This sub-section explains the machine learning algorithms used in this work. Naive Bayes (NB): This Classifier is one of the simple and most sophisticated Classification techniques helping in constructing quick ML models that can make predictions fast. As a probabilistic classifier, it offers predictions based on the likelihood that an object will occur.. The term ‘naive’ supposes that the occurrence of one attribute is unlinked to the occurrence of others’ characterstics; that is why it is called naive. E.g., if the color, taste, and shape characterize the fruit, green, sweet, and spherical are categorized as grapes. Without depending on each other, every characteristic helps recognize it as grapes.

$$P(A|B) \ = \ P(A|B)P(A) \ |P(B)$$

The term ‘Bayes’ refers to the algorithm’s dependence on Bayes’ Theorem, given above.

Logistic Regression(LR): Logistic Regression (LR) is among the most commonly used machine learning algorithm falling within the supervised learning category. With the aid of a predetermined set of independent variables, the categorized dependent variable is predicted. The categorical dependent variable's prediction is expressed as either true or false, 0 or 1, yes or no, etc. Probabilistic values lying between 0 and 1 are given by LR instead of 0 and 1. In Logistic regression, An “S” shaped logistic function is fitted rather than fitting a regression line, which presumes two largest values (0 or 1). The logistic function's curve shows the likelihood of determining whether a mouse is obese or not, as well as whether the cells are cancerous. A logistic function is a mathematical procedure for mapping estimated values to probabilities. A sigmoid function is a mathematical tool that converts the projected values to probabilities. Any actual value between 0 and 1 is changed into another value. The outcome must lie between 0 and 1. The result of logistic regression should fall in the range of 0 and 1..

K Nearest Neighbor (KNN): The KNN algorithm presupposes the newest case/data and existing cases in a comparable manner, and it places the new data point in the group closest to like that of the existing categories. After storing all of the previous data, the next data point is evaluated by using KNN algorithm similarity measures. Moreover, by deploying the KNN technique, Fresh data may be categorized quickly and accurately into the appropriate section. KNN generates zero assumptions about the data because it employs a nonparametric technique.

Support Vector Machine (SVM): To provide a solution to regression and classification problems, SVM is among the most popular supervised learning algorithm. ML classification problems are resolved by it. This algorithm aims to develop the best selection that may divide n-dimensional geometry into specific groups and also allow us to classify data points freshly and quickly. It is used to resolve problems related to Classification and Regression. It may also be employed in Machine Learning Classification issues. A hyperplane is used to explain this optimum threshold limit. It chooses points and extreme vectors to help build a hyperplane. This algorithm relies on support vectors that are used to describe extreme instances.

V. EXPERIMENT SETUP & RESULTS

We performed the experiments using four separate classifiers: SVM, LR, NB, and KNN algorithms. We combined both multilingual & monolingual datasets for the test set. We split the dataset in half, using 80% for training and 20% for testing. The classifiers are trained using the entire training set, and the results are then recorded on the test set.

TABLE III. CLASSIFIER WITH BOW VECTORIZATION TECHINIQUE

Classifier	H/NH	Accuracy	Precision	Recall	F1-Score
Logistic Regression	H NH	0.92	0.91 0.97	0.99 0.70	0.95 0.82
KNN	H NH	0.86	0.86 0.92	0.99 0.48	0.99 0.63
Naïve Bayes	H NH	0.88	0.92 0.79	0.94 0.74	0.93 0.77
SVM	H NH	0.95	0.94 0.98	0.99 0.82	0.97 0.89

TABLE IV. CLASSIFIER WITH TF-IDF VECTORIZATION TECHINIQUE

Classifier	H/NH	Accuracy	Precision	Recall	F1-Score
Logistic Regression	H NH	0.89	0.88 0.58	0.99 0.72	0.93 0.93
KNN	H NH	0.86	0.86 1.00	0.99 0.67	0.99 0.80
Naïve Bayes	H NH	0.90	0.89 0.79	0.99 0.62	0.94 0.76
SVM	H NH	0.96	0.95 0.99	0.99 0.85	0.99 0.93

The dataset’s accuracy, recall, F1-score, and precision with both vectorization techniques, namely, BoW and TF-IDF for the two classes H (Hate) and NH (Non-Hate), are shown in

Tables 3 and 4. These experiments are performed using two features, “cleantext” (created by processing the ”test case attribute) and “label gold” which are taken into account simultaneously. Among the suggested methods, the Support Vector Machine with TF-IDF records the best F1-score(0.99 for Hate contents & 0.93 for Non-Hate contents) and accuracy (0.96 for both Hate & Non-Hate). The accuracy score of 0.86 for KNN with BOW and TF-IDF vectorization is the lowest of all. KNN also has the lowest F1-score for Non-Hate content, at 0.63, with BOW. Therefore, it is clear that the optimum performance comes from a Support Vector Machine with TF-IDF vectorization.

VI. DISCUSSION AND FUTURE WORK

In this study, we used a corpus of 5,884 tweets from multilingual functionalities that had been annotated as either, profane offensive, hateful, or neither. Also discussed is the supervised approach for identifying hate speech in multilingual functionality. The corpus comprises 34 functionalities, of which six are multilingual, introduced by Das et al. [6] and the remaining 28 are based on English (monolingual), derived from Rottger et al. [22]. The source language of each word in the tweets is likewise annotated. Our classification algorithm uses features including negation words, punctuation, and target identification (hate lexicons). When utilizing SVM as the classification technique, the best accuracy of 96 percent is obtained when all the features are included in the feature vector. For future work, the corpus can be vectorized using Word2Vec and Glove, which may produce good results. These tests can also be applied in the future to code-mixed literature from multilingual civilizations that include more than two languages, together with the annotations described in this paper.

REFERENCES

[1] Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588, 2020.

[2] Michał Bilewicz and Wiktor Soral. 27 the politics of hate. The Cambridge Handbook of Political Psychology, page 429, 2022.

[3] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, pages 36–41, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[4] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media, pages 36–41, 2018.

[5] C¸ agrı C¸ ˘ oltekin. A corpus of turkish offensive language on social ¨ media. In Proceedings of the 12th language resources and evaluation conference, pages 6174–6184, 2020.

[6] Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. Hatecheckhin: Evaluating hindi hate speech detection models. arXiv preprint arXiv:2205.00328, 2022.

[7] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516, 2019.

[8] Ona De Gibert, Naiara Perez, Aitor Garc´ıa-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444, 2018.

[9] Bojan Evkoski, Andraz Pelicon, Igor Mozeti ˇ c, Nikola Ljube ˇ siˇ c, and ´ Petra Kralj Novak. Retweet communities reveal the main sources of hate speech. PloS one, 17(3):e0265602, 2022.

[10] Nimali Hettiarachchi, Ruvan Weerasinghe, and Randil Pushpanda. Detecting hate speech in social media articles in romanized sinhala. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 250–255, 2020.

[11] Gyorgy Kov ¨ acs, Pedro Alonso, and Rajkumar Saini. Challenges of ´ hate speech detection in social media. SN Computer Science, 2(2):1– 15, 2021.

[12] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Evaluating aggression identification in social media. In Proceedings of the second workshop on trolling, aggression and cyberbullying, pages 1–5, 2020.

[13] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd.

[14] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation, pages 14–17, 2019.

[15] Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schafer, Tharindu Ranas- ¨ inghe, Marcos Zampieri, Durgesh Nandini, et al. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. arXiv preprint arXiv:2112.09301, 2021.

[16] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In International Conference on Complex Networks and Their Applications, pages 928–940. Springer, 2019.

[17] Swapnanil Mukherjee and Sujit Das. Application of transformer-based language models to detect hate speech in social media. Journal of Computational and Cognitive Engineering, 2022.

[18] Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. Cold: Annotation scheme and evaluation data set for complex offensive language in english. Journal for Language Technology and Computational Linguistics, 34(1):1–28, 2020.

[19] Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in greek. arXiv preprint arXiv:2003.07459, 2020.

[20] Aneri Rana and Sonali Jha. Emotion based hate speech detection using multimodal learning. arXiv preprint arXiv:2202.06218, 2022.

[21] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. arXiv preprint arXiv:2005.04118, 2020.

[22] Paul Rottger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen ¨ Margetts, and Janet B Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. arXiv preprint arXiv:2012.15606, 2020.

[23] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. Aggressive behavior, 44(2):136–146, 2018.

[24] Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In IberLEF@ SEPLN, pages 236–245, 2020.

[25] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pages 88–93, 2016.