A New Hate Speech Detection System based on Textual and Psychological Features

Fatimah Alkomah, Sanaz Salati, Xiaogang Ma
Department of Computer Science
University of Idaho
Moscow, ID

***Abstract*—Hate speech often spreads on social media and harms individuals and the community. Machine learning models have been proposed to detect hate speech in social media; however, several issues presently limit the performance of current approaches. One challenge is the issue of having diverse comprehensions of hate speech constructs which will lead to many speech categories and different interpretations. In addition, certain language-specific features, and short text issues, such as Twitter, exacerbate the problem. Moreover, current machine learning approaches lack universality due to small datasets and the adoption of a few features of hateful speech. This paper develops and builds new feature sets based on frequencies of textual tokens and psychological characteristics. Then, the study evaluates several machine learning methods over a large dataset. Results showed that the Random Forest and BERT methods are the most valuable for detecting hate speech content. Furthermore, the most dominant features that are helpful for hate speech detection methods combine psychological features and Term-Frequency Inverse Document-Frequency (TFIDF) features. Therefore, the proposed approach could identify hate speech on social media platforms like Twitter.**

Keywords*—Hate speech detection; hate speech classification; hate speech features; hate speech methods

I. INTRODUCTION

As the number of users of social media increases, the impact of hate speech is drastic due to the ease of posting hate speech without geographical boundaries and user anonymity. The uncontrolled spread of hate can damage our society gravely and severely harm marginalized people or groups [1]. The effect of hate crimes is widely spread due to the users' anonymity[2] and the wide use of social media. Twitter, as social media, was studied by 54.81% of researchers; primarily, textual analysis was the prevalent method with 33% compared to other methods [3].

Hate speech detection is a challenging research problem due to many issues, including competing definitions, limited feature sets, small-sized datasets, and the current design of current models. Competing hate speech definitions capture different information with different interpretations by proposed models. For example, racist and homophobic tweets are more likely to be classified as hate speech. However, some definitions are debatable [4]. Therefore, the nonexistence of a universally accepted definition is due to whether offensive conveys hate or not [5]. The critical aspect is separating hate speech language from other offensive languages [6]. The problem of competing definitions would result in a poor feature detection set that could not help identifying hate speech. The problem posed by ungrammatical text has mainly been used to mitigate the difficulty of automatically detecting hateful speech, particularly when users intentionally change keywords' spelling or avoid automatic content [7], [8].

The issue of feature detection becomes more challenging as some words are contextual dependent on users and groups and are not inherently offensive [9], [10]. Small-sized datasets are not enough to generalize results or capture compelling hate speech detection features. For example, Cervero's method [11] employs 200 tweets and yet achieves a good result. Obstacles also include partially labeled data, which makes comparing the performance of many datasets hard to validate. Therefore, many machine learning models do not generalize any hate speech content as it is limited to specific keywords or dictionaries [11]. For example, it was shown that the Yin and Zubiaga model's performance[12] drops down by 10% when tested on another dataset outside the same group of datasets. As a result, the feature sets of datasets do not necessarily represent real-life cases, despite reported performance[11]. Therefore, several machine learning models cannot scale well in practice or models that are not robust due to dataset bias.

This paper develops several machine learning models that are helpful in detecting hate speech based on textual tweets on Twitter. The paper uses the Twitter dataset of 150k tweets [13], called the MMHS150K benchmark dataset. The images were removed from the dataset, and the dataset was converted from the JSON to a tabular format. Three textual features were extracted from the dataset: the frequency of user mentions, hashtags, and emojis; TFIDF of 3-grams; and psychological features extracted by the Linguistic Inquiry and Word Count (LIWC) [14]. Linguistic Inquiry and Word Count is a software application for counting words that references a lexicon of grammatical, psychological, and content word categories. LIWC has been used to categorize texts effectively along psychological dimensions (such as users' personality traits and emotions). The proposed approach was tested on Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, and Decision Trees algorithms. This study aims to present a model that could be used to automate hate speech detection on any social media platform such as Twitter. We also aim to find the best features that work well with the best-performing algorithm.

The proposed method has several contributions aside from using existing machine learning models from conventional and deep learning methods. This study has extensively studied

the effect of three different groups of feature sets on the results of hate speech detection. We have shown that combining more than one feature set provides a good performance model. Moreover, the proposed method studies the multilabel classification problem and delivers results at the label level, which was lacking in previous studies. Additionally, the proposed model could be integrated with social media platforms to instantly detect and block hate speech.

Our research objectives include identifying textual features that were effective in the classification. For example, the model should be able to detect hate speech fine-grained at the label level given a short text (Tweet).

The paper is outlined as follows. Related works are summarized in Section II. Section III illustrates the proposed machine learning approach. Results and discussions are explained in Section IV and V. The paper is concluded in Section VI.

II. Related Work

Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) is a new track to detect hate speech detection in the research community. The HASOC track intends to provide a platform to develop and optimize Hate Speech detection algorithms for Hindi, German and English [15]. The best result on the English language dataset of HASOC was based on Long Short-Term Memory (LSTM), which used GloVe embeddings as input. The best system achieved a performance of f1-measure of 0.52; however, the dataset has only 3,708 records for the English dataset. The International Workshop on Semantic Evaluation (SemEval) organizes the OffensEval series of shared tasks on offensive language identification using the hierarchical annotation of the type and target of offensive content [16] [17]. However, robust datasets survive in many classification tasks of hate speech and are reusable and easy to update.

Furthermore, it was reported that robust datasets are required to allow comparability of features and methods [18]. Therefore, as posts of hate speech can also be implicit, few lexical features could be used for machine learning models. Although there are many approaches and features, the current list of models cannot be generalized due to dataset size, credibility, low precision, or imbalanced datasets.

The literature reported various features of hate speech that include shallow lexical features [19], dictionaries [20], sentiment analysis [21], linguistic characteristics [22], knowledge-based features [23], and meta-information [24] of social media content. Readers may refer to a comprehensive study of hate speech detection methods and datasets published recently [25]. However, the literature showed that shallow lexical detection methods have low precision [19]. The literature reported that identifying hate speech on a large scale is still an unsolved problem [26]. For example, the DeepHate method [16] is based on many features: word embeddings, sentiment, and topic information. Recently, aggressive and gendered identification are getting attention [27]. It was found that stylometric (such as function words) and emotion-based features are robust indicators of hate speech[28]. Markov *et al.* [28] provided a model based on encoded emotion information of 14,182 emotion words and their association with emotions and sentiments from the emotion lexicon [29]. Furthermore, the Linguistic Inquiry and Word Count (LIWC) of Pennebaker *et al.* [14] and profanity [30] (especially anger) are good indicators of hate speech in the Indian language context [31]. The LIWC categories include linguistic statistics such as counts and summary variables: analytic, clout, authenticity, and emotional tone. In addition, the LIWC could reveal feelings, personality, and psychological motivations [14]. However, it was shown that the features relating to users' personality traits and emotions in text achieved an accuracy result of 0.7 in English text [32].

Therefore, current methods lack a suitable set of features for hate speech; are either based on small datasets or have low performance when tested over multiclassification hate speech problems. The overall issue is related to the nonexistence of a universally accepted definition of hate speech which results in whether offensive tweets convey hate or not [5].

III. Proposed Framework

In this study, the proposed framework is a machine learning model with an input of a hate speech dataset and trained binary classification output. The framework (Fig. 1) has four steps: data preparation, feature extraction, model learning, and classification output.

A. Data Preparation

It was found that datasets target multiple hate speech categories; however, only 60% of dataset builders reported an inter-annotator agreement [33]. Moreover, it is common for many datasets to overlap between class labels, as Waseem [34] showed an overlap of 2,876 tweets with the Waseem and Hovy datasets [35]. Therefore, relevant and no obsolete datasets are essential to a useful predictive hate speech model. However, creating large and varied hate or abusive datasets that minimize potential bias is laborious and requires specialized experts [36]. Therefore, this study uses a large benchmark dataset taken from a previous Twitter dataset of 150k tweets [13], the MMHS150K dataset. The dataset has an average tweet length of 91 characters, a minimal length of 15, and a maximum length of 193, including the URLs. The dataset has images and textual data of tweets and image captions from Twitter in a python dictionary inside a JSON file. The key of each entry in the JSON file is the tweet ID. The other fields include three different fields, which are the image URL, tweet URL, tweet text, and class labels. The dataset has six classes, shown in Table. I.

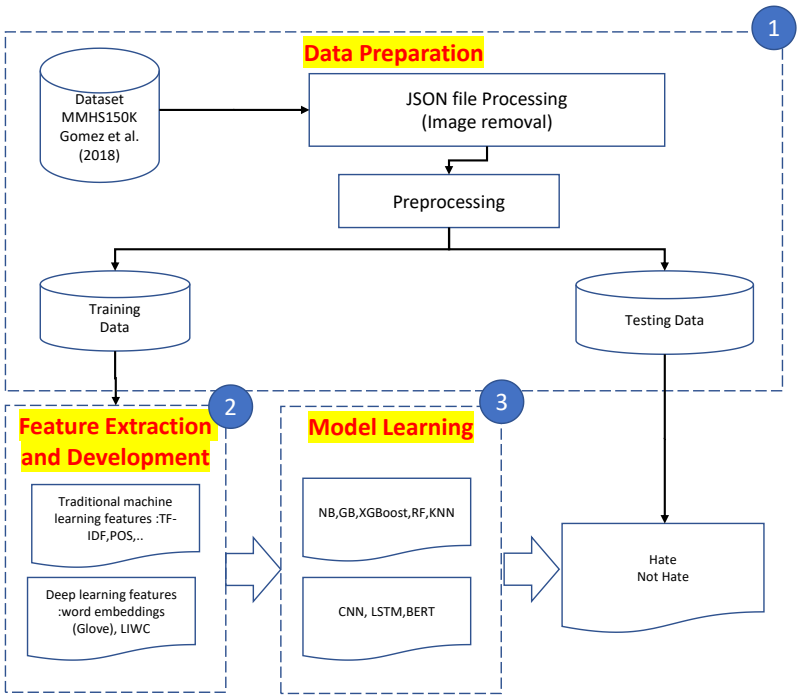


Fig. 1. Proposed Framework.

TABLE I. DISTRIBUTION OF CLASS LABELS IN THE BENCHMARK DATASET*

Class	Total Instances
Not Hate	131,081
Racist	44,535
Sexist	19,509
Homophobe	10,554
Religion	2,119
Other Hate	21,217
Total	143,277

* Further details about extracting and preparing the original dataset are found here https://gombru.github.io/2019/10/09/MMHS/

B. Data Preprocessing

The following are the text preprocessing actions carried out in this study.

1) Removal of images and keeping only textual content in the dataset. This step involves converting the dataset into a tabular format for further preprocessing.

2) Stop words removal.

3) Convert text to lowercase after counting the number of capital letter words.

4) Removal of user mention after checking if a tweet has a mentioned user.

5) Emotions extraction using the UNICODE_EMOJI library from the emot.emo_unicode package.

6) Convert emojis to placeholders so that they will be part of the 3-grams.

7) Tokenization.

8) Lemmatization.

9) 3-grams Extraction.

*10)*Convert text to TF-IDF vector.

C. Feature Extraction and Development

Based on previous literature, this study selects several feature sets such as frequency of tokens (e.g., hashtags) or TFIDF and word embeddings. We follow the following criteria for selecting the sets of features: (1) features must be used in prior hate speech detection models with evidence of acceptable results, (2) the feature must be textual and in line with the current dataset characteristics, and (3) the feature should be used by at least two related studies. Therefore, following these criteria, the features are explained in Table II.

Notably, the selection of feature set 3 is used by only one related study; however, such feature set (LIWC) was evident in other studies related to human sentiments. Therefore, different combinations of the three groups will be used with various machine learning algorithms.

D. Model Learning

This study examines the performance of traditional and deep learning methods on the benchmark dataset. A good model must use the minimum number of features; therefore, this study finds the best features that maximize performance. Consequently, the following methods were selected from machine learning: Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, and Decision Trees. The benchmark dataset was split into training and testing (80% training and 20% for testing). Stratified sampling is used to ensure proper sampling for each class label. The dataset is imbalanced; therefore, the dataset is balanced using oversampling techniques of SMOTE, where BorderlineSMOTE was the best.

TABLE II. THE FEATURE SET FOR HATE SPEECH DETECTION

Category	Name	Description	Rationale	Related Studies
Feature Set 1 (Counts)	Username Mention	Checks if a tweet mentioned any other user. The preprocessing uses '1' if a username is mentioned, '0' otherwise	Mentioning a person could indicate hate toward that person.	[37]
	Capital Letter	Count of words with Capitals letters.	Tweets that have capital letters may indicate hate or stress in speech.	[38]
	Hashtags	Count of hashtags	Similar to the user mention	[37] [39] [40]
	Emojis	Count of emojis in the tweets	Could indicate a negative attitude of users	[26]
Feature Set 2 (TFIDF)	TFIDF Or Word Embeddings	Cleaned TFIDF vector. It includes tokenization, stemming, and 3-grams. Word embeddings (Glove) is used with Deep learning models.	The semantic structure of tweets	[41] [42] [37] [43] [44] [45]
Feature Set 3 (LIWC)	Summary Language Variables	analytic, clout, authentic, and tone	Analytical thinking, clout, authenticity, emotional tone	Proposed by this study
	Linguistic Dimensions	i, we, you, shehe, they, and ipron	It was shown that the hate speech features relating to users' personality traits and emotions in text achieved an accuracy result of 0.7 in English text [32].	Proposed by this study
	Psychological Processes	Affect variables posemo, negemo, anx, anger, and sad		[46]
	social	family, friend, female, and male		Proposed by this study
	bio	body, health, sexual, and ingest		Proposed by this study
	drives	affiliation, achieve, power, reward, and risk		Proposed by this study

The results that predict hate speech are anlayzed following the standards of machine learning. The precision, recall, f1-measure, and ROC standard performance metrics. The analysis also includes a comparison with other methods in hate speech detection. The analysis also includes identifying the most predictable features for each model.

IV. EXPERIMENTS, RESULTS AND DISCUSSION

Following the selected features in Table II and after preprocessing the dataset, each feature set was created using the python scikit-learn library. Each feature set was prepared alone, allowing different feature sets to be combined with various machine learning algorithms. For example, the first feature sets were processed as follows:

1) If a Tweet includes a usernames, add another feature that is '0' has does not include a username and and '1' has mentioned any user name.

2) If the Tweet has capital letter words, add a new feature and place the total number of capitalized words.

3) Keep hashtags as they might be informative.

4) Replacing URLs, mentions, emojis, Retweets, and capital letters as placeholders.

For example, the tweet

'@Mr Rodie94 Nigga was in the store like 😂 https://t.co/dSEb83kIhm'.

becomes

'mention placeholder nigga was in the store like face_with_tears_of_joy url_placeholder'.

The second feature set is the TFIDF with 3-grams, which was carried out using python; the top trigrams are shown in Fig. 2. Samples of preprocessing steps are shown in Fig. 3. Finally, the third feature set was extracted with LIWC software, which in turn was exported to excel and preprocessed with python.

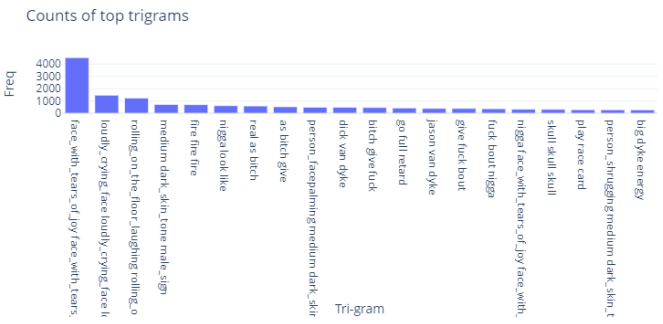


Fig. 2. Top Trigrams.

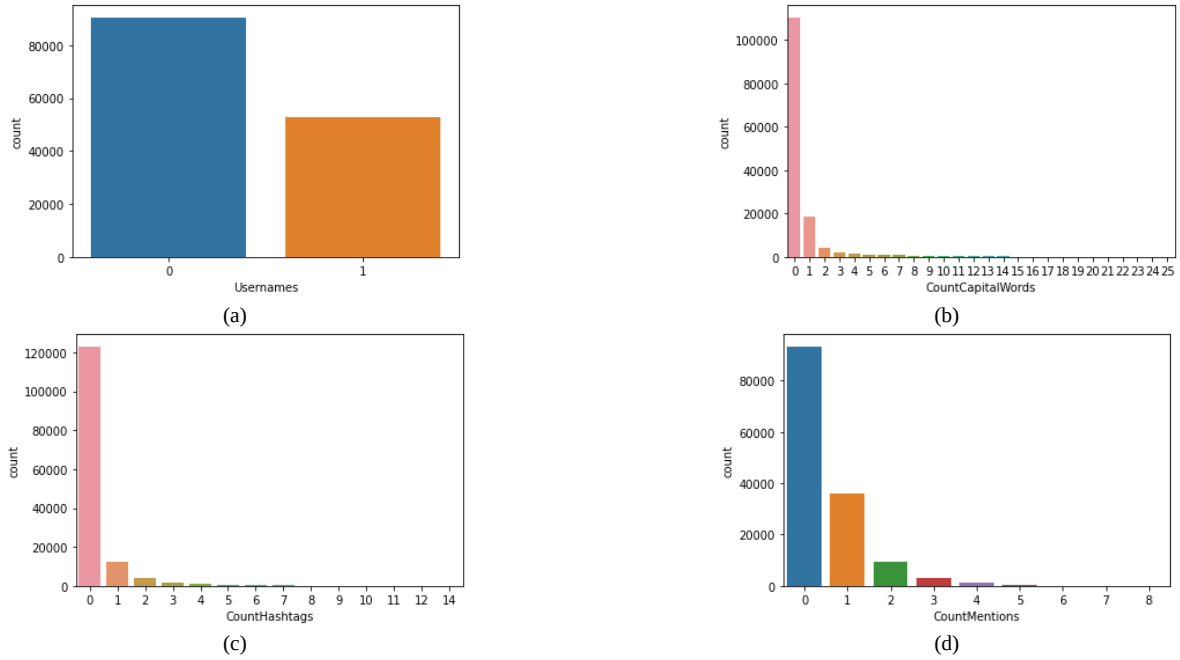


Fig. 3. Preprocessing Steps. Showing Usernames for Both Negative and Positive Example (a), Counting of Capital Words (b), Counting of Hashtags(c), and Count of Mentioned (d).

E. Application of the Proposed Methods (Model Learning)

The classification of this research is a binary classification where each machine learning algorithm is tested on the dataset (hate/not hate). The adopted methods are explained in Table III. On the other hand, the deep learning structure for binary classification of hate speech is shown in Appendix A. The parameters were deduced as per many experiments considering that the nature of machine learning is multiclassification. Each feature set was first to run alone with a specific method, and then the features were combined together.

TABLE III. TRADITIONAL MACHINE LEARNING METHODS FOR BINARY CLASSIFICATION OF HATE SPEECH

Algorithm	**Settings**
Naïve Bayes	naive_classifier = MultinomialNB()
Gradient Boosting	criterion='friedman_mse', init=None, learning_rate=0.1, loss='deviance', max_depth=3, max_features='log2', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=80, n_iter_no_change=None, random_state=None, subsample=1.0, tol=0.0001,
	validation_fraction=0.1
XGBoost	base_score=0.5, booster='gbtree', colsample_bytree=0.6, gamma=0.3, learning_rate=0.01, max_depth=3, min_child_weight=1, n_estimators=20, random_state=40, reg_alpha=0, reg_lambda=1.5, scale_pos_weight=1, seed=None, subsample=0.4
Random Forest	n_estimators = 200
KNN	algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=10, p=2, weights='uniform'
Decision Trees	riterion='entropy', random_state=1
Deep Learning models	
CNN	Structure varies based on feature sets, as explained in A ppendix A
LSTM	Structure varies based on feature sets, as explained in A ppendix A
BERT	Structure varies based on feature sets, as explained in A ppendix A

F. Classification Output Analysis

Following the machine learning Table III, Appendix A, and the proposed set of features in Table II, the results are depicted in Fig. 4-6 and discussed here. As shown in Fig. 4, the first feature set is the lowest-performing feature set, indicating that such features are not performing well. However, the second and the third feature sets provide promising results with the most studied algorithms. The highest performance was for the BERT, with a 0.974 f1-score measure on the second feature set and 0.956 on both the first and the third feature sets. The f1-measure for positive and negative examples of the selected machine learning models is shown in Fig. 5. The figure shows that models provide high performance for positive examples (hate=1) and low performance for negative examples. This finding is consistent with previous works [19] and shows that negative examples are still challenging due to the use of similar keywords, as illustrated earlier [6] [5]. Therefore, the nonexistence of a universally accepted definition is due to whether offensive conveys hate or not [5]. Overall, the proposed model provided higher performance in binary classification, 0.98 compared to the original model of a maximum of 0.734 [47].

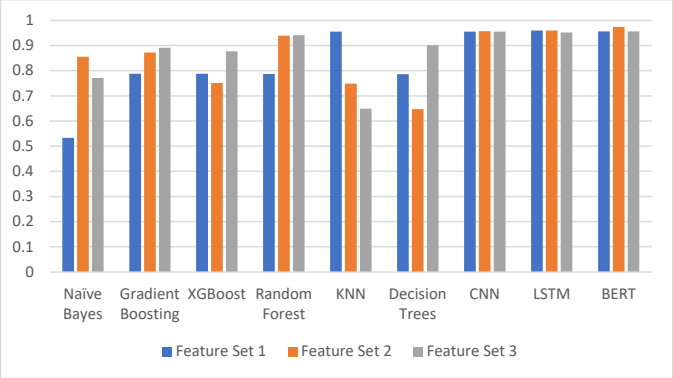


Fig. 4. Traditional and Deep Machine Learning Algorithms F1-Measure against Feature Sets.

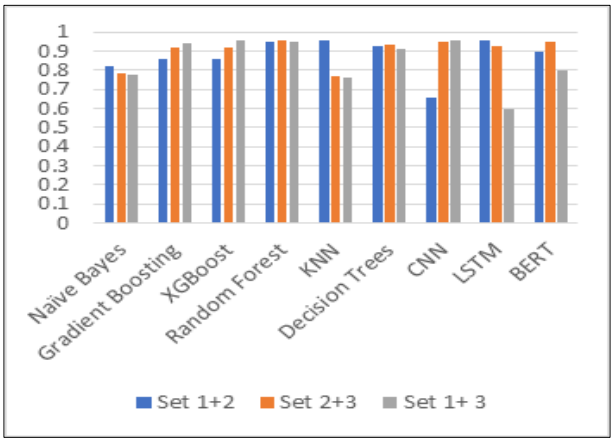


Fig. 5. Selected Algorithms Average F1-Measure (Binary Classification Feature Combinations).

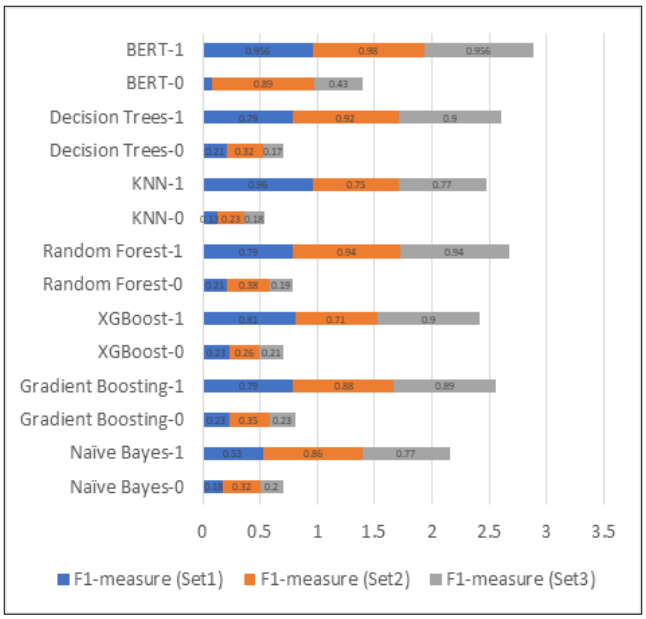


Fig. 6. Deep Learning Machine Learning Algorithms F1-Measure against Feature Sets.

Next, the performance of LSTM, CNN, and BERT (along with the baseline methods) are shown in Figure 6. For BERT: bert_multi_cased_L-12_H-768_A-12/2 model were used. The f1-measure for the BERT model is the highest among the deep learning models. The structure of these algorithms is shown in Appendix A. As compared with previous methods, BERT is the most promising method. The reported f1-measure for BERT is 0.974. Above all, BERT was the most prominent method that distinguishes the negative examples of hate speech, as shown in Fig. 5. However, in practice, it is essential to select the best performing set of features that provides the optimal model. Fig. 5 shows the list of selected algorithms and their performance when several features are merged. It shows that the LSTM got an f1-measure of 0.96 when combining features (set 1 and set 2). Contrary to our previous finding that BERT is the best, it was not performing as compared to LSTM due to the complexity of integrating feature sets of the original bert_multi_cased model and the new features extracted from text. Nevertheless, the most consistent algorithm for the random forest provides relatively similar results when different feature sets.

Table IV shows a sample of related works on hate speech classification and their performance. Unfortunately, most models are not available to the public and were tested in different datasets. Therefore, careful interpretation of the results in the table should be considered as different datasets will eventually change the model outcomes; this issue is already discussed before.

TABLE IV. TRADITIONAL MACHINE LEARNING METHODS FOR BINARY CLASSIFICATION OF HATE SPEECH

Ref	Dataset	Best Method	Accuracy	F1-measure
[48]	Islamophobic hate speech:100K tweets	One-versus-one SVM	0.77	
[19]	25K tweets	SVM		**0.91**
[49]	5K tweets	Logistic Regression	0.704	
[39]	76 K tweets	MCD + LSTM	0.78	
[50]	6.6K tweets	GRU + CNN		**0.78**
[51]	14 K for SemEval-2019 Task 6 subtask A: Offensive/non-offensive	MCD + LSTM	0.78 F1-score	
[52]	SemEval-2019 Task 6 [154]	GRU + CNN	Task A: classification of tweets into either offensive (OFF) or not offensive (NOT) 0.78 for supervised 0.77 for the unsupervised approach	
[53]	**six datasets and 121 customized list**	**Cat Boost**		**F1-score ranging from 0.85 to 0.89 Best average F1-score 87.74 across all datasets**

V. DISCUSSION

Our research objectives include identifying textual features that were effective in the classification. The research showed that the most dominant features are textual features extracted from TFIDF features, as shown in Fig. 2. The features are focused on emotional features such as face_with_tears_of_joy, which was evident in the dataset with 4,528 frequent items. In addition, other keywords were frequent, such as 'fire,' 'nigaal', 'dick van dyke', and others. Such a finding is consistent with previous studies that showed that sentiments are effective in showing a large number of hate speech contents [37], [38], [54]. In addition, the findings are consistent with works related to LIWC as additional features showing human behavior [46].

The developed machine learning models showed that, as expected, the binary classification was providing acceptable results. The best performing model was BERT with 0.974. LSTM also reported good results with an f1-measure of 0.96. The reason is that these models depend on high-dimensional word embedding, and their design was proved to work well with many textual classification tasks. The combination of feature set 2 and feature set 3 provides good results for LSTM and BERT models. The other models reported lower performance, such as CNN (below 0.66 f1-meaure) for the combination of feature set 1 and feature set 2. A single feature set, such as feature set 2 performed well on most algorithms. The best-performing model reported an f1-score of binary classification f1-score of 0.704 with the Feature Concatenation Model (FCM) [13]. The proposed model reported LSTM with an f1-measure of 0.96 (feature set1+feature set 2) with binary classification and 0.96 on LSM and CNN (feature set 2+feature set 3). However, the proposed model has not reported good performance for each label. The investigations showed the original imbalanced dataset, which does not have enough examples for each label. Due to the complexity of hate speech detection, decision trees and KNN provided high f1-measure performance based on TFIDF feature sets. However, these algorithms did not generalize well at the label level (hate/not hate), indicating that there were standard features between positive and negative examples of the hate speech benchmark dataset.

Consequently, with a wide set of machine learning models, the results indicate that as the number and type of features are added (shown in groups in Table II.), the machine learning model performance increase. The reason is that the additional features add new semantics to the embedded or intended meaning in a particular Tweet. For example, the LIWC features (Feature set 3), have shown relatively good performance in detecting sentiments and user psychological features.

Although the experiments have been run on a single dataset, the dataset is considered one of the largest datasets that are available online. According to a previous study, it was found that current datasets suffer from various aspects, including their size, bias, and authenticity in terms of the annotation process [25]. A comparison of hate speech models was not fully available as many models are not published, or the dataset is private. However, the proposed model was able to provide an acceptable accuracy with a baseline work that used additional non-textual features such as images and their captions [13]. Therefore, given these restrictions, and due to the complexity of hate speech features, the results are considered acceptable but should be interpreted within the context of hate speech categories implied in the adopted benchmark dataset. The new work provides implications to theory with newly adapted machine learning models and could be used on unseen data on Twitter or similar social media platforms.

VI. CONCLUSION

This paper develops three feature sets that could be used for hate speech detection: frequencies of unique tokens, TFIDF, and LIWC features. Then, the paper extensively compares several machine learning models: Naïve Bayes, Gradient Boosting, XGBoost, Random Forest, KNN, Decision Trees, LSTM, CNN, and BERT. The difficulty of hate speech identification was shown by the high f1-measure performance of decision trees and KNN based on TFIDF feature sets. However, these algorithms did not generalize effectively at the label level (hate/not hate), showing that positive and negative samples of the hate speech benchmark dataset shared common

characteristics. Conversely, the results of the BERT model were relatively higher, with an f1-measure of 0.974 on the same feature set (TFIDF). In addition, the LIWC feature sets and their combination with TFIDF provided better results on the LSTM method. However, features among the adopted LIWC could share common information. It is recommended that the adopted approach should be considered in the context of generic hate speech on a short text like Twitter. The model might need retraining due to out-of-vocabulary keywords that users might use over time. Furthermore, the researchers might consider another resource of hate speech aside from Twitter. Therefore, we plan to test the models based on a single sub feature on a leave-out scheme in the future.

REFERENCES

[1] M. Bedrosova, H. Machackova, J. Šerek, D. Smahel, and C. Blaya, "The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech Republic, Poland, and Slovakia," Comput. Human Behav., vol. 126, p. 107013, 2022, doi: https://doi.org/10.1016/j.chb.2021.107013.

[2] S. T. Peddinti, K. W. Ross, and J. Cappos, "User anonymity on twitter," IEEE Secur. Priv., vol. 15, no. 3, pp. 84–87, 2017.

[3] A. Matamoros-Fernández and J. Farkas, "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," Telev. New Media, vol. 22, no. 2, pp. 205–224, 2021, doi: 10.1177/1527476420982230.

[4] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions," Comput. Sci. Rev., vol. 38, p. 100311, 2020, doi: 10.1016/j.cosrev.2020.100311.

[5] N. Strossen, "Freedom of speech and equality: Do we have to choose," JL Pol'y, vol. 25, p. 185, 2016.

[6] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLoS One, vol. 14, no. 8, pp. 1–16, 2019, doi: 10.1371/journal.pone.0221152.

[7] A. Nourbakhsh, F. Vermeer, G. Wiltvank, and R. van der Goot, "sthruggle at SemEval-2019 Task 5: An Ensemble Approach to Hate Speech Detection," pp. 484–488, 2019, doi: 10.18653/v1/s19-2086.

[8] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," SN Comput. Sci., vol. 2, no. 2, pp. 1–15, 2021, doi: 10.1007/s42979-021-00457-3.

[9] S. Ullmann and M. Tomalin, "Quarantining online hate speech: technical and ethical perspectives," Ethics Inf. Technol., vol. 22, no. 1, pp. 69–80, 2020, doi: 10.1007/s10676-019-09516-z.

[10] E. Mosca, M. Wich, and G. Groh, "Understanding and Interpreting the Impact of User Context in Hate Speech Detection," no. Ml, pp. 91–102, 2021, doi: 10.18653/v1/2021.socialnlp-1.8.

[11] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying generalisability across abusive language detection datasets," in Proceedings of the 23rd conference on computational natural language learning (CoNLL), 2019, pp. 940–950.

[12] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: a review on obstacles and solutions," PeerJ Comput. Sci., vol. 7, pp. 1–38, 2021, doi: 10.7717/PEERJ-CS.598.

[13] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020, pp. 1459–1467, 2020, doi: 10.1109/WACV45572.2020.9093414.

[14] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[15] T. Mandl et al., "Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages," CEUR Workshop Proc., vol. 2826, pp. 87–111, 2020.

[16] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 1415–1420, 2019, doi: 10.18653/v1/n19-1144.

[17] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification," Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021, pp. 915–928, 2021, doi: 10.18653/v1/2021.findings-acl.80.

[18] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," no. 2012, pp. 1–10, 2017, doi: 10.18653/v1/w17-1101.

[19] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017, no. Icwsm, pp. 512–515, 2017.

[20] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis," Behav. Inf. Technol., vol. 39, no. 7, pp. 711–721, 2020, doi: 10.1080/0144929X.2019.1607903.

[21] F. H. A. Shibly, U. Sharma, and H. M. M. Naleer, Classifying and Measuring Hate Speech in Twitter Using Topic Classifier of Sentiment Analysis, vol. 1165. Springer Singapore, 2021. doi: 10.1007/978-981-15-5113-0_54.

[22] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," 12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018, no. ICWSM, pp. 42–51, 2018.

[23] H. Abburi, S. Sehgal, and H. Maheshwari, "Knowledge-based Neural Framework for Sexism Detection and Classification," no. September, 2021.

[24] F. Rangel, G. L. D. L. P. Sarracén, Bert. Chulvi, E. Fersini, and P. Rosso, "Profiling Hate Speech Spreaders on Twitter Task at PAN 2021," CLEF 2021 Labs Work. Noteb. Pap., no. September, pp. 21–24, 2021.

[25] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Information, vol. 13, no. 6, 2022, doi: 10.3390/info13060273.

[26] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A Multilingual Evaluation for Online Hate Speech Detection," ACM Trans. Internet Technol., vol. 20, no. 2, 2020, doi: 10.1145/3377323.

[27] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Evaluating Aggression Identification in Social Media," Proc. Second Work. Trolling, Aggress. Cyberbullying, no. May, pp. 1–5, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.1.

[28] I. Markov, N. Ljubesic, D. Fiser, and Walter, "Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection," Proc. 11th Work. Comput. Approaches to Subj. Sentim. Soc. Media Anal., pp. 149–159, 2021, [Online]. Available: https://www.aclweb.org/anthology/2021.wassa-1.16/.

[29] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word--emotion association lexicon," Comput. Intell., vol. 29, no. 3, pp. 436–465, 2013.

[30] T. Jay and K. Janschewitz, "The pragmatics of swearing," 2008.

[31] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, "Did you offend me? Classification of Offensive Tweets in Hinglish Language," pp. 138–148, 2019, doi: 10.18653/v1/w18-5118.

[32] R. Cervero, "Use of Lexical and Psycho-Emotional Information to Detect Hate Speech Spreaders on Twitter," 2021.

[33] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," Lang. Resour. Eval., vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.

[34] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," pp. 138–142, 2016, doi: 10.18653/v1/w16-5618.

[35] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," pp. 88–93, 2016, doi: 10.18653/v1/n16-2013.

[36] B. Vidgen and L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, vol. 15, no. 12 December. 2021. doi: 10.1371/journal.pone.0243300.

[37] Z. Ziqi, D. Robinson, and T. Jonathan, "Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network," IJCCS (Indonesian J.

Comput. Cybern. Syst., vol. 11816 LNAI, no. 1, pp. 2546–2553, 2019, [Online]. Available: https://doi.org/10.475/123_4.

[38] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[39] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with hindi and english social media," Inf., vol. 12, no. 1, pp. 1–16, 2021, doi: 10.3390/info12010005.

[40] S. Masud et al., "Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter," Proc. - Int. Conf. Data Eng., vol. 2021-April, pp. 504–515, 2021, doi: 10.1109/ICDE51399.2021.00050.

[41] C. M. V. de Andrade and M. A. Gonçalves, "Profiling Hate Speech Spreaders on Twitter: Exploiting textual analysis of tweets and combinations of multiple textual representations," CEUR Workshop Proc., vol. 2936, pp. 2186–2192, 2021.

[42] E. Ombui, L. Muchemi, and P. Wagacha, "Hate Speech Detection in Code-switched Text Messages," 3rd Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2019 - Proc., pp. 1–6, 2019, doi: 10.1109/ISMSIT.2019.8932845.

[43] A. Joulin et al., "Deep Learning for Hate Speech Detection in Tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, vol. 2, no. 2, pp. 759–760. doi: 10.1145/3041021.3054223.

[44] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text Analysis for Hate Speech Detection Using Backpropagation Neural Network," Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018, pp. 159–165, 2018, doi: 10.1109/ICCEREC.2018.8712109.

[45] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the Arabic language context," ICPRAM 2020 - Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods, no. January, pp. 453–460, 2020, doi: 10.5220/0008954004530460.

[46] N. Bauwelinck, G. Jacobs, V. Hoste, and E. Lefever, "LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval)," pp. 436–440, 2019, doi: 10.18653/v1/s19-2077.

[47] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020, pp. 1459–1467,. doi: 10.1109/WACV45572.2020.9093414.

[48] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," J. Inf. Technol. Polit., vol. 17, no. 1, pp. 66–78, Jan. 2020, doi: 10.1080/19331681.2019.1702607.

[49] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," EPJ Data Sci., vol. 5, no. 1, 2016, doi: 10.1140/epjds/s13688-016-0072-6.

[50] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," no. 7491, pp. 85–90, 2017, doi: 10.18653/v1/w17-3013.

[51] S. Modha, P. Majumder, and D. Patel, "DA-LD-Hildesheim at SemEval-2019 Task 6: Tracking Offensive Content with Deep Learning using Shallow Representation," pp. 577–581, 2019, doi: 10.18653/v1/s19-2103.

[52] G. Wiedemann, E. Ruppert, and C. Biemann, "UHH-LT at SemEval-2019 Task 6: Supervised vs. Unsupervised Transfer Learning for Offensive Language Detection," pp. 782–787, 2019, doi: 10.18653/v1/s19-2137.

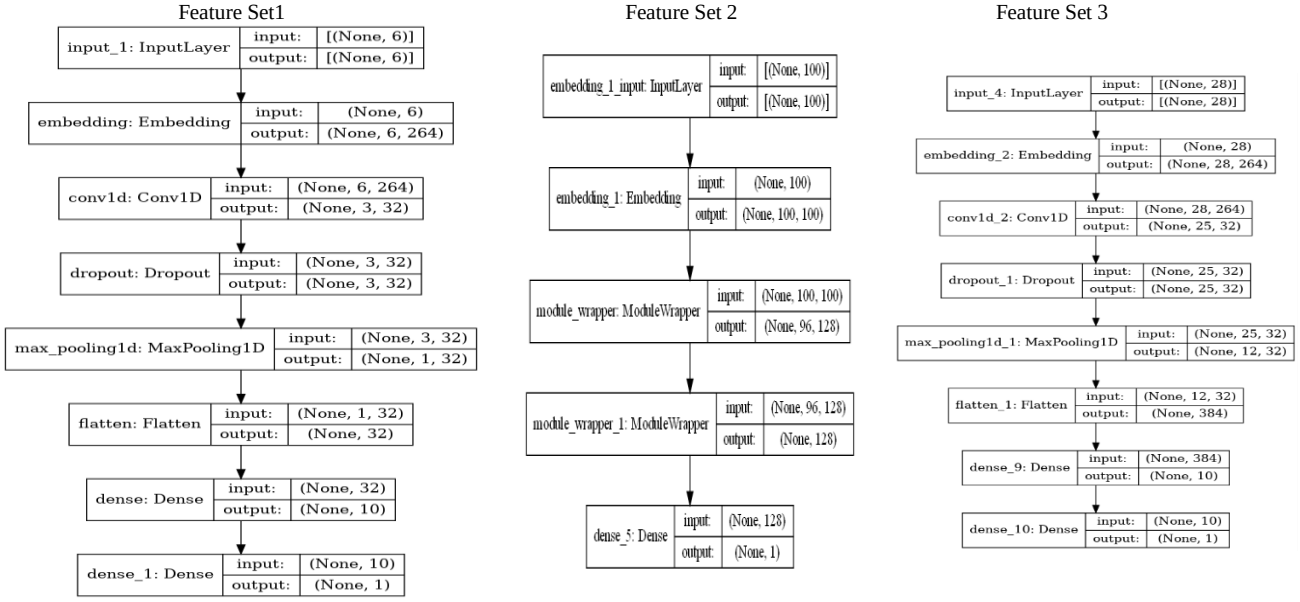
[53] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," IEEE Access, vol. 9, pp. 109465–109477, 2021, doi: 10.1109/ACCESS.2021.3101977.

[54] A. T. E. Capozzi et al., "Computational linguistics against hate: Hate speech detection and visualization on social media in the 'Contro L'Odio' project," CEUR Workshop Proc., vol. 2481, pp. 0–5, 2019.

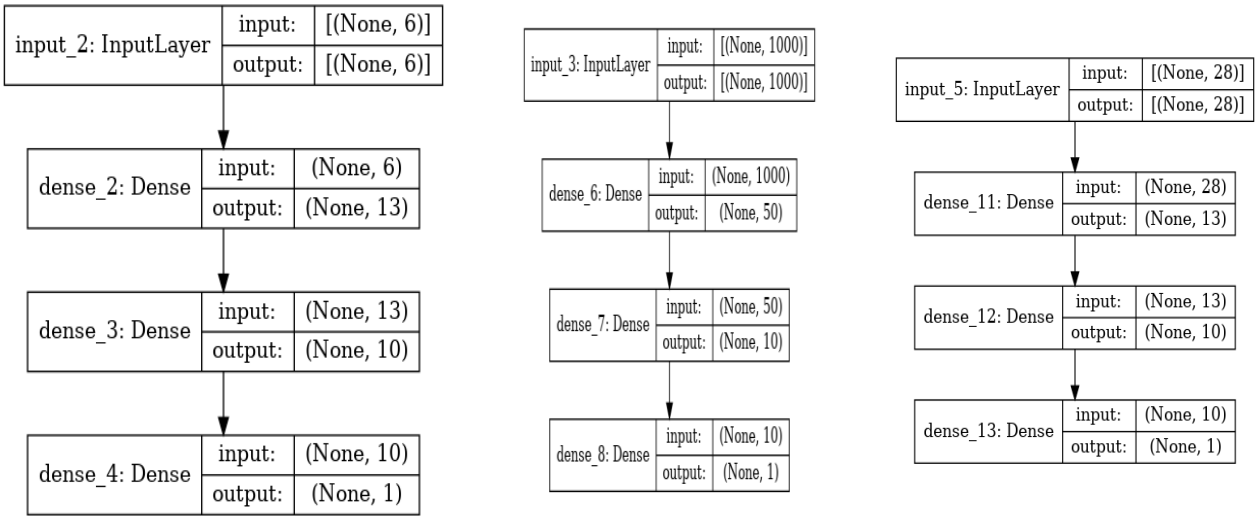
APPENDIX

Appendix A: Architecture of deep learning models used in this paper. Please note that the parameters of these algorithms was tuned based on 10% of the dataset after several trails on the algorithms till the parameters were set as shown the appendix.

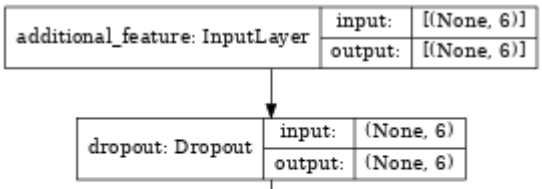
CNN



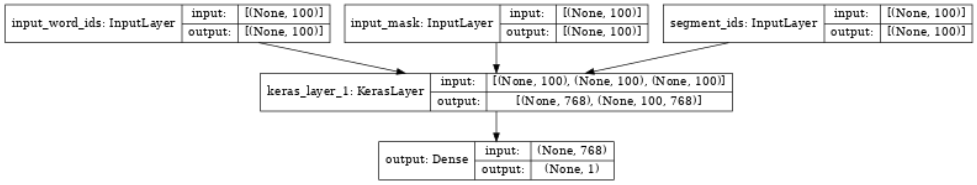
LSTM



BERT
Feature set 1



BERT
Feature set 2



BERT
Feature set 3