Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media

Yasas Senarath
George Mason University
Fairfax, Virginia, United States
ywijesu@gmu.edu

Hemant Purohit
George Mason University
Fairfax, Virginia, United States
hpurohit@gmu.edu

***Abstract*—Detecting malicious intent behavior such as sharing hate speech has become an important challenge for social networking platforms. The method of automated hate speech detection for social media posts is often challenged by the complexity of capturing the context of the user expression with potential hate intent. We hypothesize that semantic features can help enrich the context representation of word senses in a social media post for machine learning algorithms. This paper presents a novel empirical study of diverse semantic features for hate speech classification task on social media posts. Specifically, we present an extensive empirical analysis, where we test the features of vector space model representation for corpus-based semantics, neural word embedding representation for distributional semantics, and declarative knowledge patterns from external knowledge base for domain semantics.**

Our experimental results show that ensembling the diverse feature representations improves the efficiency of hateful behavior classification in contrast to the case of a single type of feature representation. Results on two popular Twitter datasets for the hate speech detection task showed a consistent performance gain for the classification models that were based on the hybrid feature representation (absolute gain in $F1$ score up to 3.0%). The application of the proposed method of combining diverse feature representations can help in improving social media analytics systems for monitoring human behavior.

***Keywords*-Semantic Text Classification, Hate Speech, Social Media, Malicious Intent, Feature Engineering**

I. Introduction

Social media has become an integral part of our daily activities. This makes the proliferation of offensive and hateful behavior in social media platforms a significant concern for our society. Such behavior has grave implications for individual and societal levels ranging from the polarizing and incivil conversations to mental health issues.

Platforms such as Facebook and Twitter employ several human users to monitor and manually filter the offensive and hateful content, however, it is a very time-consuming process. Researchers, therefore, have explored automated techniques, primarily supervised classification methods [1], [2] that are trained on social media posts labeled by human annotators. However, there are several challenges in efficiently detecting hateful behavior with these automated techniques.

First challenge, in fact, is to specifically define the behavior of hateful intent in a social media post, due to the subtle nature of user expression in a given context that can mislead someone for interpretation, e.g., sarcasm vs. angry rant vs. hate speech. This challenge has led researchers to study different types of fine-grained behaviors in the social media content such as offensive, abusive, hate [1]–[4], cyberbullying [5], and aggression [6]. The second challenge is the ability of the detection algorithms to better formalize and represent the context of the user expression in a given social media post. Thus, prior studies have explored multiple feature types to improve context representation for the hate speech detection task, however, with less exploration for the knowledge base features capturing data semantics [1]. There is no extensive exploration of how different feature representations and their combinations corresponding to diverse semantic information in a given text play a role in the complex task of natural language understanding for hate detection.

TABLE I
EXEMPLAR POSTS ON SOCIAL MEDIA WITH HATEFUL INTENT. *(Messages slightly rephrased for anonymity)*

	Message	**Characteristics**
M1	members of nontraditional religions r all sub-human trash	*hateful*
M2	you sure u ain't colored?	*offensive/ hateful*
M3	such a sucker 4some Oreos.	*not hateful*

Table I shows some example posts for offensive and hateful behavior. *M1* is a prototypical hateful message threatening some religious communities explicitly. *M2* is offensive but implicitly hateful given the context of color used for the race, however, there is ambiguity. Finally, *M3* is not a hateful message but just an angry expression of a user.

Problem. We address the problem of binary classification for offensive or hateful content vs. normal content detection on social media.

Our contribution. This study presents an extensive evaluation of the significance of diverse semantic feature representations of social media messages for the complex task of hate speech detection from natural language. We introduce different semantic features in Section IV, followed by several experimental schemes for classification models with different features in Section V. We discuss the results for classification performance on two real-world Twitter datasets and implications for future work in section VI.

TABLE II
SUMMARY OF DATASETS AND THE DISTRIBUTION OF *hate /offensive speech* VS. *normal* LABELS FOR TWITTER POSTS.

Tweet Dataset	Hate/Offense	Normal	Total
DWMW17	20620	4163	24783
FDCL18	8587	51640	60227

II. RELATED WORK

There are several studies in recent years on the subject of hate speech detection on social media, due to the emergent problem of its implications causing societal polarization. We summarize the related works on automated hate detection methods for social media.

We can categorize the previous studies related to hate speech detection into different granularity levels. The least granular approach comprises of detecting whether a given social media post exhibits a hateful nature [7]. However, recently there have been many studies on multi-class hate classification. Those research works use algorithms to identify different types of hate in social media, e.g., Founta et al. (2018) [4] conducted such a study on multiple, distinct hate classes: offensive, abusive, hateful speech, aggressive, and cyberbullying. Although multi-class hate speech detection provides more insight, due to the class imbalance and sparsity problems, the performance is often very poor for the relevant class of hate speech, leading many researchers to address the binary classification task.

The computational methods for hate detection mainly focus on supervised learning algorithms [1]. Previous studies have experimented with a diverse set of techniques from conventional machine learning and state-of-the-art deep learning algorithms. Popular techniques include Naive Bayes [8], [9], Logistic Regression [3], [9], Random Forest [3], [8], etc. Although Support Vector Machine (SVM) classifier is one of the most commonly used in hate speech classification studies in the past [7]–[10] and thus, we resolved to experiment with this technique in our study.

Given the dependence of machine learning algorithms on features to make the prediction, researchers investigated different features. Multiple previous studies confirm the predictive power of surface-level, corpus-based word features for hate speech detection. According to [1], [3], other feature types used for classification are word generalization, sentiment analysis, lexical resources, linguistic features, knowledge-based features, meta-information, and multi-model information. Although such features are used rigorously in the literature, the usage of knowledge-based features is limited. Also, the state-of-the-art pre-trained language models provide additional neural embedding representations of words that can be utilized.

III. DATASETS: OFFENSIVE AND HATE SPEECH DATASETS

In this study, we have used two popular offense and hate speech datasets to train and evaluate our models. We modified the datasets to satisfy our requirements of binary classification. Table II summarizes the distribution of labels in our dataset.

DWMW17 [3] - consists of around 25k tweets collected by querying for words in hatebase.org (a lexicon for hate words). The dataset is annotated with labels: Hate, Offensive and Neither. Given the potential hate intent behind the hate and offensive speech categories, we have combined them as a single *hate* behavior category.

FDCL18 [4] - includes around 60k tweets randomly sampled from Twitter stream (recrawled using tweet-ids). The tweets are labeled with four classes: Normal, Spam, Abusive and Hateful. For our study, we have combined Spam & Normal tweets as *normal* category and Abusive & Hateful tweets as *hate* behavior category given the malicious user intent.

IV. IMPLEMENTING THE OFFENSIVE AND HATE SPEECH DETECTION MODEL

The proposed method for offensive and hate speech detection has two main steps: a.) Feature Extraction and, b.) Classification Model.

Feature Extraction. We represent each tweet (T) in the corpus as a feature vector using each of the following feature extraction methods:

- *Corpus-based semantic features* (BoW): Each tweet (T) is pre-processed using the following steps: normalizing special tweet objects (URLs and twitter mention indicator) with special tokens, tokenizing, and removing stopwords. Each pre-processed T is then converted to a vector of *tf-idf* features. We used Tweet Tokenizer in the NLTK library in python. We also allow n-grams in the range [1, 3] to appear in the resulting feature vector.
- *Declarative knowledge-based semantic features*. These provide the sense interpretation of the words in a natural language content from the human-engineered external knowledge bases.
  - *Hatebase features*: Hatebase is a structured knowledge base available online on multilingual hate speech (http://hatebase.org). Hatebase provides knowledge of hate-related terms, e.g., the definition of a hate word as well as its multiple hate-related meanings and non-hate-related meanings as well. The following list contains several word-level feature vectors (*H_x*) that are used in our approach to generate knowledge-based features of a given T, by averaging over such vectors of all Hatebase words present in T:
    a.) Offensiveness (*H_offensiveness*): a numerical score representing the offensiveness of a given word, represented as a feature vector of discrete bins. Bin-edges and the number of bins are automatically calculated using Freedman Diaconis Estimator [11].
    b.) Unambiguous (*H_unambiguous*): a boolean feature indicating whether or not a given word has an unambiguous meaning in a language.
    c.) Hateful-Meaning (*H_hateful*): all hateful definitions of hate words are used to create the vocabulary of bag-of-words model, where the presence of words in the hateful definition is used in the bag-of-words vector.

d.) Non-hateful-Meaning (*H_nonhateful*): all non-hateful definitions of hate words are used to create the vocabulary of bag-of-words model, where the presence of words in the non-hateful definition is used in the bag-of-words vector.

– *FrameNet features* (FN): FrameNet is a knowledge base that provides a rich linguistic resource of textual examples with similar latent meanings under the semantic frame categories. The tweet *T* is processed through frame semantics parsing tool *SLING* [12], which outputs PropBank frames [13] that we map to FrameNet frames using an existing method [13]. The resulting list of FrameNet frames constitutes a vocabulary to construct a vector for *T* indicating the frequency count of each frame observed in *T*.

- *Distributional semantics-based features* (E_mean): The mean of pre-trained word embeddings of the words in *T* is computed as feature vector, providing a generalized sense representation of words learned from external data. We used 300-dimensional `word2vec` embeddings [14].

Classification Model. Our classification model is based on SVM algorithm, which is trained on the input data described in Section III. We used an SVM implementation available in the scikit-learn library in python with a linear kernel and 'l2' penalty of 1.0. We used the same parameters across all datasets and features to be consistent with the results.

V. Experimental Setup and Evaluation

We conduct a robust evaluation of the various classification model schemes for offensive and hate speech detection by combining the diverse set of semantic features described earlier:

- **[*M1*] - BoW** (baseline): This scheme includes only the *tf–idf* representation of the pre-processed tweet as features.
- **[*M2*] - M1 + H_offensiveness**: The scheme includes features from M1 concatenated with the average offensiveness feature from Hatebase.
- **[*M3*] - M2 + H_unambiguous**: This scheme combines features of M2 scheme with the feature based on "is_unambiguous" metadata from Hatebase.
- **[*M4*] - M3 + H_hateful**: In addition to the features from scheme M3, this scheme uses average of bag-of-word vectors of meaning of hate words from Hatebase as described in section IV.
- **[*M5*] - M4 + H_nonhateful**: This scheme concatenates features of scheme M4 and "H_nonhateful" features of Hatebase.
- **[M6] - M5 + FN**: In scheme M6, we concatenate FrameNet features ("FN") with features in scheme M5.
- **[M7] - M6 + E_mean**: Scheme M7 is obtained by concatenating features from M6 with mean embedding features.

Evaluation metrics. To compare the different features we use common measures from machine learning literature: accuracy, precision, recall, and F1, averaged across each fold of the 5-fold cross validation setting, for each dataset. In the experiments, stratified folds are made to preserve the distribution of each class in the dataset.

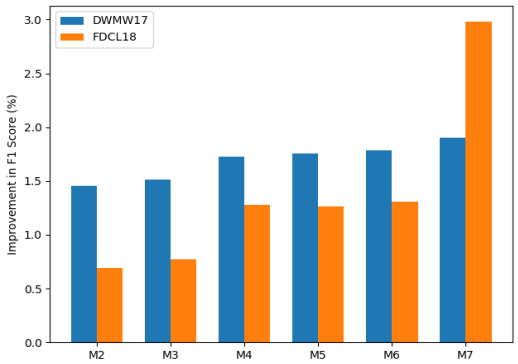

Fig. 1. Absolute gain in F1 score of each modeling scheme against the baseline (M1) for datasets DWMW17 and FDCL18.

VI. Results and Discussion

TABLE III
PERFORMANCE EVALUATION OF FEATURES AND THEIR COMBINATIONS USING 5-FOLD CROSS VALIDATION FOR EACH HATE SPEECH DATASET, WHERE BOLD VALUES INDICATE THE BEST PERFORMING SCHEME.
(Abbreviations: Mi - Model scheme identifier, A - Accuracy, P - Precision, R - Recall, and F1 - F1 Score)

Model	DWMW17				FDCL18			
	A	P	R	F1	A	P	R	F1
M7	**94.8**	97.1	**96.7**	**96.9**	**94.6**	90.0	**70.1**	**78.8**
M6	94.6	**97.2**	96.4	96.8	94.4	90.8	67.7	77.5
M5	94.6	97.1	96.3	96.7	94.4	90.8	67.6	77.5
M4	94.5	97.1	96.3	96.7	94.4	90.9	67.6	77.5
M3	94.2	97.1	95.9	96.5	94.3	90.8	67.1	77.1
M2	94.1	97.0	95.9	96.4	94.3	90.7	67.0	77.1
M1	91.6	93.5	**96.7**	95.1	94.2	**91.2**	66.0	76.5

TABLE IV
CROSS-DATASET PREDICTION PERFORMANCE OF THE BASELINE (M1) AND BEST (M7) MODELS FOR ASSESSING THEIR GENERALIZABILITY.
(Abbreviations: Mi - Model scheme identifier, A - Accuracy, P - Precision, R - Recall, and F1 - F1 Score)

TRAIN	DWMW17				FDCL18			
TEST	FDCL18				DWMW17			
Model	A	P	R	F1	A	P	R	F1
M7	**83.1**	**44.7**	77.9	**56.8**	**82.7**	97.8	**81.0**	**88.6**
M1	62.7	25.7	**85.7**	39.6	70.8	**98.7**	65.7	78.9

Result observations. Table III compares the performance of classification models with different feature combinations. We note the following insights:

- ***tf-idf* features capturing corpus-based semantics are highly predictive, although for a given dataset.** These features provide good classification performance (above 90% accuracy in Table III) when the model is trained and tested on the same event dataset; however, as shown in Table IV, they do not help generalize the models when applied across dataset domains.

TABLE V
EXAMPLES WHERE THE HYBRID FEATURE MODEL M7 DETECTS (✓) OR MISSES (X) IN CONTRAST TO THE BASELINE M1 THAT MISSED ALL.

Tweet	Label	Baseline	Best Model
person still questions my love for Oreos lol	FALSE	X	✓
No these are only whites pies. Well do u have any negro pies?	TRUE	X	✓
momma said no pussy cats inside my doghouse	FALSE	X	X
Happy birthday to my pops. Old negro spiritual is 64 today lol	FALSE	X	X

- **Declarative knowledge base features enhance precision.** Knowledge base features such as Hatebase features and FrameNet features help more precisely predict the hate speech due to enhanced contextual sense representation of the hate words. For instance, in Table III for dataset DWMW17, this is evident from comparing the precision and recall for M1 and M6 modeling schemes.
- **Distributed semantic features also help improve performance.** Figure 1 shows the modeling scheme M7 based on word embedding features with all other knowledge base features improves F1 scores. Since external knowledge bases may provide features for some words, the pre-trained word embeddings help enable sense interpretation for larger vocabulary of words, helping the classification. Also, a big increment in F1 score for FDCL18 dataset could be attributed to the imbalanced composition of the dataset.

Error Analysis. Table V shows examples and the related original labels (*TRUE* indicating hate tweet and *FALSE* indicating Normal tweet) as well as whether the label is detected by the baseline model and the hybrid model scheme M7. We note that our proposed modeling scheme perform better for some cases, it also fails on some instances although the baseline model fails there too. It suggests the need for further enhancement of contextual interpretation; for instance, in the 4th tweet, understanding the relationship between the potential receiver and the author of the tweet might help (hint 'pops'). We plan to explore such discourse characteristics in the future study.

Limitations and Future Work. While the features introduced in this paper improve the performance of hate speech detection, it has several limitations that guide towards future work. *Polysemy* words with multiple meanings can hinder the actual text interpretation, presenting a challenge to detect the sense of hate intent. Although, this challenge was partially addressed by the introduced knowledge base features, future work needs to contextually disambiguate the senses of words. Also, we only explored the hypothesis for introducing semantic features to detect hate intent in English language text; thus, future studies can explore multilingual social media posts.

VII. CONCLUSION

This paper presents a novel empirical study of diverse semantic feature representations for hate speech detection on social media. We showed that the semantic features can help enrich the context representation of word senses for machine learning algorithms. We demonstrated the applicability of the hybrid feature representation approach for efficient hateful behavior detection that can provide complementary contextual information derived for a given content, as shown by the results (absolute gain in $F1$ score up to 3.0% for the models with hybrid feature representation). The application of our method can help in improving data analytics systems for social media streams across a variety of application domains, including public safety, governance, and journalism.

Reproducibility. The code for experiments is available at https://git.gmu.edu/ysenarath/public/hate-intent-detection.

Acknowledgement. We thank US National Science Foundation grant IIS-1657379 for partial research support.

REFERENCES

[1] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. of the Fifth Int'l Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10.

[2] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," in *SemEval*, 2019, pp. 75–86.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *ICWSM*, 2017.

[4] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *ICWSM*, 2018.

[5] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Tran. on Interactive Intelligent Systems*, vol. 2, no. 3, p. 18, 2012.

[6] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proc. of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018, pp. 1–11.

[7] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[8] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *PASSAT-SOCIALCOM*. IEEE, 2012, pp. 71–80.

[9] Y. Mehdad and J. Tetreault, "Do characters abuse more than words?" in *Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 299–303.

[10] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *CIKM*. ACM, 2012, pp. 1980–1984.

[11] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L 2 theory," *Probability theory and related fields*, vol. 57, no. 4, pp. 453–476, 1981.

[12] M. Ringgaard, R. Gupta, and F. C. Pereira, "Sling: A framework for frame semantic parsing," *arXiv preprint arXiv:1710.07032*, 2017.

[13] M. Palmer, "Semlink: Linking propbank, verbnet and framenet," in *Proc. of the generative lexicon conference*. GenLex-09, Pisa, Italy, 2009, pp. 9–15.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.