LING 573 Deliverable #2 Report

Mohamed Elkamhawy
University of Washington
mohame@uw.edu

Karl Haraldsson
University of Washington
kharalds@uw.edu

Alex Maris
University of Washington
alexmar@uw.edu

Nora Miao
University of Washington
norah98@uw.edu

Abstract

The paper describes our preliminary affect recognition system developed for SemEval-2019 Task 5, which focuses on identifying hate speech in tweets targeted at immigrants and women. The task is divided into two subtasks: the primary task involves hate speech detection in English tweets, while the adaptation task addresses Spanish tweets. We employed a binary classification approach, in which each given tweet is classified as either hate speech or non-hate speech. To build our system, we trained a Word2Vec model to generate word embeddings and then used them as input features for a Support Vector Machine (SVM) classification algorithm. Our current system represents an initial attempt at creating a functioning hate speech detection model, and it lays the foundation for future improvements and adaptations.

1 Introduction

In recent years, the proliferation of hate speech on social media platforms has become a critical societal issue receiving unprecedented attention. Not only does hate speech have detrimental effects on individuals' psychological well-being, but it also poses significant safety risks to our society at large (Saha et al., 2019). Hate speech can lead to devastating implications. For example, it can perpetuate and reinforce racism, biases, stereotypes, and discrimination against certain individuals, eliciting emotional distress. Moreover, studies have shown a link between hate speech and increased acts of violence and hate crimes against members of minority groups (Relia et al.). As online social networks continue to reach broader audiences and serve as a primary means of communication, the need for more effective hate speech detection algorithms has become particularly evident. Recent research in automatic hate speech detection using NLP techniques has shown some promising results[1]. To address the spread of hateful narratives online, we built a preliminary model that aims to identify hate speech on Twitter against two specific minority groups: immigrants and women.

The rest of the paper is organized as follows. Section 2 provides an overview of the primary and adaptation tasks that we aim to address, along with a description of the dataset and the evaluation procedure. Section 3 presents the overall system architecture and briefly discusses the major design decisions. Next, Section 4 details the major subcomponents of the system. Section 5 outlines the key findings obtained through the evaluation of the system, while Section 6 interprets the findings and analyzes the results in detail. Additionally, this section includes error analysis and assessment of each component. Furthermore, in Section 7, we discuss the limitations and ethical considerations regarding the implementation and application of our system. Finally, Section 8 concludes the paper by summarizing the main findings and potential next steps for improvements.

2 Task Description

The chosen primary task is detecting hate speech in tweets, where the hate speech is against either women or immigrants, as described in SemEval-2019 Task 5 (Basile et al., 2019). Specifically, it is a binary classification task targeted at determining attitude–here whether a given tweet contains hate speech or not. The genre for this task is tweets and the modality is text. The target of this task is aspect-specific, and the language is English for the primary task. The adaptation task will be hate speech identification in Spanish tweets, also described in Basile et al. (2019). The only difference between the primary task and the adaptation task is

[1] See, e.g., Asogwa et al. (2022), Kotarcic et al. (2023), and Schmidt and Wiegand (2017).

the language, while all the other dimensions remain unchanged.

The data for the shared task was collected from July to September 2018 for the immigrant-targeted tweets (Basile et al., 2019). The data for the women-targeted tweets was collected from July to November 2017 (Fersini et al., 2018). The English language dataset contains 13,000 tweets, 9,000 of which are in the training set, 1,000 in the development set, and 3,000 in the test set. Of the 13,000 tweets, 7,530 are annotated for the negative class, and 5,470 are annotated for the positive class. The Spanish language dataset that will be used for the adaptation task consists of 6,600 tweets, 4,500 of which are for training, 500 for development, and 1,600 for testing. Of the 6,600 tweets, 3,861 are annotated for the negative class, and 2,739 are annotated for the positive class. The annotations were collected using the *Figure Eight* (F8) platform, where each tweet was annotated by at least three contributors, and then a relative majority label was assigned. Expert annotators were also utilized, such that the final label of a given tweet was determined by the majority label of the F8 annotation and two independent expert annotators (Basile et al., 2019). The evaluation is calculated using accuracy, precision, recall, and macro-averaged F1-scores to maintain class-size independence since the hate speech and non-hate speech class sizes are relatively balanced (Basile et al., 2019).

The dataset can be requested using this form, and the shared task is detailed at this page as well as in Basile et al. (2019).

3 System Overview

We developed an end-to-end system using `Python3` for the primary task. Our development workflow consists of six stages: data preprocessing, feature extraction, training, inference, evaluation, and deployment. In the data preprocessing stage, we cleaned and prepared data using the Natural Language Toolkit (NLTK) library[2], and performed standard text processing such as removing special characters and stopwords, tokenization, and lemmatization. In the feature extraction stage, we employed a Bag-of-Words approach for our baseline model. For our proposed models, we generated word embeddings using a pre-trained word2vec model from the Gensim library, which transformed text data into numerical representations. Additionally, we utilized the `Empath` package in conjunction with `Word2Vec` to generate lexical features for the second proposed model.

Once the features were prepared, we used a support vector machine (or SVM, a supervised learning model) as our classifier and trained the model using labeled data in the training set. The inference stage involves using the trained SVM model to predict the presence of hate speech in tweets from the development dataset. In the evaluation stage, we assessed the performance of our model using the evaluation metrics prescribed in Basile et al. (2019), including standard metrics such as Accuracy, Precision, Recall, and F1 score. Lastly, we integrated all the components into a comprehensive, functional system that can be run in the patas environment.

4 Approach

The system implements a series of binary classifiers for detecting hate speech in tweets. Our approach comprises four primary components: (1) data preprocessing; (2) feature extraction; (3) model training; (4) model inference; and (5) evaluation.

We evaluate three models. The baseline model relies on a "bag of words" representation for the features and a binary support vector machine (SVM) classifier (the model hereafter known as BOW-SVM). Our first proposed model also uses a binary SVM classifier but relies on embeddings derived using `word2vec` (hereafter W2V-SVM). Our second proposed model uses additional lexical features derived using the `Empath` package (hereafter W2V-EMP-SVM).

4.1 Data preprocessing

The Shared Task data is provided in raw CSV format, already split into train, dev, and test sets. Each row in the CSV files represents a Tweet. The CSV includes a column with a numerical unique identifier (id), a column with the text of the tweet (text), and three target variable columns (HS, TR, and AG). Column HS is the target for Subtask A, and therefore this system. In that column, the value 1 indicates that hate speech is present. The value 0 indicates that it is not.

This system ingests that data, then performs

[2]Natural Language Toolkit (NLTK) is a popular open-source library for natural language processing (NLP) in Python. It provides a suite of tools for text-processing tasks such as classification, tokenization, stemming, tagging, and parsing. More information can be found at `https://www.nltk.org/`.

several text preprocessing operations to ready the data for modeling. First, the removes hashtags and splits the text of the hashtags into distinct tokens using capitalization. Second, it removes all URLs and characters related to mentions (e.g., "@"). Third, it tokenizes the text using a specialized tweet tokenizer from NLTK, which takes into account tweet-specific elements including hashtags, mentions, and emoticons. Fourth, it lemmatizes these tokens using the WordNet lemmatizer[3] available from NLTK. Finally, it removes English stopwords[4] as defined in NLTK.

These steps are applied consistently for every modeling approach, including the baseline BOW-SVM.

4.2 Feature extraction

Our system applies two distinct feature engineering approaches: one for the baseline model and another for the two models we developed specifically for this task. For the baseline model, we implement a simple BOW mechanism via `scikit-learn`'s `CountVectorizer()` object. For the proposed models, we rely on `word2vec` (continuous bag of words). W2V-EMP-SVM includes the additional step of concatenating on the `word2vec` representation matrix a vector of lexical information for each instance, using the `Empath` package. `Empath` is a tool that is used to score text across 194 pre-identified categories that have been determined from topics and emotions, using dependency relationships from the ConceptNet knowledge base, and compiled seed terms that correspond to each concept. [5].

4.3 Model training

Each of BOW-SVM, W2V-SVM, and W2V-EMP-SVM uses an SVM classifier. We apply to our training set a basic grid search approach and 5-fold cross-validation to identify the best hyperparameters.[6] Each model is scored based on its macro-averaged f1 score and the best model is saved and used for evaluation.

4.4 Model inference

Our system accepts at the time of inference both models trained during the same run and models saved from a prior run. In either case, the system applies the model to the preprocessed and engineered tweet text and outputs predictions for each dev or test instance.

4.5 Evaluation

The system applies the evaluation measures for Subtask A as described in Basile et al. (2019). Those measures are accuracy, precision, recall, and F1-score. Submissions are ranked by macro-averaged F1-score (Basile et al., 2019). As prescribed by the Shared Task, we calculate each using the `precision_recall_fscore_support()` method from `sklearn.metrics`. Those metrics are then written to a results file.

5 Results

Our system's results on the development dataset for the shared task are presented in Table 1.

Table 1: SemEval-2019 Task 5 (Subtask A) Results

Model	Acc	Prec	Rec	F1-Macro
W2V-SVM	0.67	0.62	0.71	0.60
W2V-EMP-SVM	0.70	0.66	0.72	0.65
BOW-SVM *Baseline*	**0.73**	**0.72**	**0.73**	**0.72**
Basile SVC *Baseline*	-	-	-	0.45

Basile et al. (2019) provides a baseline of a support vector classifier (SVC) that uses default hyperparameters and a tf-idf approach. When evaluated on the test set, the F1-macro score for that model is 0.45. We've included it here for reference, but it is worth noting that our D2 results are based on performance against true labels in the development set, not the test set. We, therefore, establish a baseline model of our own as well. It implements a bag-of-words approach with a cross-validated support vector machine classifier (BOW-SVM), which achieved an accuracy of 0.73, a precision of 0.72, a recall of 0.73, and an F1-Macro score of 0.72.

Our SVM models that rely on word2vec representations had lower scores for each. The W2V-SVM model achieved an accuracy of 0.67, precision of 0.62, recall of 0.71, and an F1-Macro score

[3] More information about the WordNet lemmatizer can be found at `https://www.nltk.org/api/nltk.stem.wordnet.html#nltk.stem.WordNetLemmatizer`

[4] Stopwords are a set of commonly used words in a language that does not add any additional information or meaning to a sentence. They are generally filtered out during the data-preprocessing phase of many NLP tasks. The NLTK corpus provides a list of stopwords for multiple languages. More information can be found at `https://www.nltk.org/book/ch02.html#stopwords_index_term`.

[5] See Fast et al. (2016) for more details

[6] Searching across all combinations of C {0.1, 1, 10} and kernel {linear, rbf, sigmoid}, the parameters with the highest F1-macro score were a C of 0.1 and the linear kernel

of 0.60. Similarly, the W2V-EMP-SVM model achieved an accuracy of 0.70, precision of 0.66, recall of 0.72, and an F1-macro score of 0.65. These models have lower F1-macro scores than our own baseline BOW-SVM model, but they were more than 0.15 points over the baseline in Basile et al. (2019).

6 Discussion

Error analysis was conducted on the pre-processed dev tweets to identify if there were any attributes in common among the false negatives and false positives for each of the approaches, and across approaches, in the overlapping mislabeled documents as well. The dev data contains 573 tweets labeled for the negative class and 427 tweets labeled for the positive class.

Table 2: Errors on Development Set

Model	TP	TN	FP	FN
BOW-SVM	281	452	121	146
W2V-SVM	131	537	36	296
W2V-EMP-SVM	174	522	51	253

The W2V errors highly favor the false negatives, where 89.2% and 83.2% of the errors are false negatives for W2V-SVM and W2V-EMP-SVM respectively. Of the false positive errors, (BOW-SVM, W2V-SVM) have 26 errors in common, (BOW-SVM, W2V-EMP-SVM) have 35 errors in common, and (W2V-SVM, W2V-EMP-SVM) have 33 errors in common. Of the false negative errors, (BOW-SVM, W2V-SVM) have 135 errors in common, (BOW-SVM, W2V-EMP-SVM) have 130 errors in common, and (W2V-SVM, W2V-EMP-SVM) have 249 errors in common. Therefore, the approaches, especially the two W2V approaches and for false negatives, seem to be making mostly the same errors.

Table 3: Classwise Model Performance on Hateful Tweets Only

Model	Precision	Recall	F1
BOW-SVM	0.70	**0.66**	**0.68**
W2V-SVM	**0.78**	0.31	0.44
W2V-EMP-SVM	0.77	0.41	0.53

When we inspect the classwise precision and recall of the BOW-SVM model and the W2V-SVM model, we see that the BOW-SVM model recalls 66% of hateful tweets compared to just 41% for the W2V-EMP-SVM model. One reason for this may be that information present in a simple BOW approach is lost when embeddings are applied. That is, applying and averaging embeddings may have eliminated important signals from specific tokens. We expect that changes to the preprocessing approach and, potentially, to a more recent contextual representation (e.g., using word-piece tokenization and a transformer-based model), might help recover some of the information loss.

Specific text features are observed frequently in mislabeled tweets. For the BOW-SVM approach, out of the 67 tweets that contain the word 'skank,' the model classified 32 (47.8%) of those tweets as positive for hate speech, while the true count is 47 (70.1%). It is likely that the feature 'skank' was not identified as a potential indicator of a tweet containing hate speech. On the other hand, out of 13 tweets that contained the word 'money,' the model classified 11 (84.6%) of those tweets as positive for hate speech, while the true count is 6 (46.2%). The model may have erroneously identified the feature 'money' as a signal for hate speech. For example, the following tweet is a false positive, which may be at least somewhat implicitly derogatory, and mentions immigrants (specifically "send back mexico" in the pre-processed tweet):

> **Original**:@AmericaNewsroom Illegal Immigrant's don't want to be Citizen's. They want the money exchange. A US dollar in Mexico is worth Ten dollars. If you get Three dollars a hour and room and board then send it back to Mexico it is really Thirty dollars a hour.
>
> **Pre-processed**: illegal immigrant want citizen want money exchange u dollar mexico worth ten dollar get three dollar hour room board send back mexico really thirty dollar hour

For the W2V-SVM approach, out of 121 tweets that contain the word 'hysterical,' the model classified 0 (0%) as positive for hate speech, while the true positive count is 65 (53.7%). The model correlated the feature 'hysterical' with non-hate speech, when in fact the tweets that contain that word are not necessarily so. Notably, 62 out of the 65 true positive tweets also contain the word 'woman,' and for all the tweets containing 'woman,' the model classified them as containing hate speech

only 2.2% of the time (5 out of 227 tweets). The below example illustrates a false negative:

> **Original**: @SenGillibrand You have lost all sense of reality. You're one very HYSTERICAL woman.
>
> **Pre-processed**: lost sense reality one hysterical woman

Similarly, for the W2V-EMP-SVM approach, 9 (7.4%) of the 'hysterical' tweets were classified by the model as positive for hate speech. Interestingly, out of 38 tweets containing 'shut,' the model classified 29 (76.3%) as positive for hate speech while the true count is 20 (52.6%); and out of the 23 tweets containing 'love,' the model classified 4 (17.4%) as positive for hate speech while the true count is 11 (47.8%). This indicates that the moral loadings of those words were correlated to hate speech being identified accordingly, i.e. 'shut' is correlated with being positive for hate speech (and in the 38 tweets containing 'shut', 33 also contain 'fuck,' which is the highest co-occurring word with 'shut'), and 'love' with negative.

Further, when calculating the average empath scores over all lexical categories, the category of violence had the largest difference in average scores between true positives and false positives, where the true positives averaged 0.0038 and the false positives averaged 0.0092. The category of childish had the largest difference in average scores between true negatives and false negatives, where the true negatives averaged 0.0044 and the false negatives averaged 0.0123. For reference, the largest scoring categories for each of true positives, false positives, true negatives, and false negatives respectively are swearing terms (0.0427), swearing terms (0.0458), crime (0.0104), and childish (0.0123).

7 Ethical Considerations

Our preliminary model has several limitations. Firstly, hate speech can change over time as language and culture evolve. As previously mentioned, the training data we used was collected in 2018. Therefore, our model may not be able to adequately capture the recent forms of hate speech as it lacks contextual understanding of the current world. Moreover, our existing system may not be particularly adept at identifying implicit hate speech, which is often expressed through sarcasm, metaphor, or other subtle forms of expression. Detecting implicit hate speech can be a challenging task as it requires context and relies less on lexical cues. As a result, our system needs to be constantly updated and fine-tuned in order to adapt to the nuanced and ever-changing patterns of hate speech.

The implementation and application of hate speech detection models against immigrants and women raise several ethical challenges that need to be considered. Firstly, the data was collected using a keyword-driven approach (Basile et al., 2019), which may lead to a biased and unrepresentative dataset. Furthermore, privacy is another ethical concern that is worth noting. The raw tweets from the dataset contain mentions and other information, which could potentially be used to re-identify the user who made the post and preferred to stay anonymous. Hate speech detection models may also amplify biases towards minority groups. For instance, certain terms may be considered offensive when used by a certain group of people but not necessarily by others. Hate speech depends on social and cultural context. To mitigate the potential harms associated with the deployment of hate speech detection models, it is imperative to build a system that is transparent, responsible, and accurate.

8 Conclusion

Our baseline BOW-SVM model achieved the best overall performance among the three models that we evaluated. Although our W2V-SVM model and W2V-EMP-SVM model did not perform better than our baseline model, they still achieved reasonable results and improved upon the Basile et al. (2019) baseline SVC model. To further improve our model performance, we will experiment with different preprocessing techniques, explore transformer-based models, and fine-tune hyperparameters on the development set.

References

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification using SVM and Naive BAYES. *arXiv preprint arXiv:2204.07057*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Min-

neapolis, Minnesota, USA. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 214–228.

Ana Kotarcic, Dominik Hangartner, Fabrizio Gilardi, Selina Kurer, and Karsten Donnay. 2023. Human-in-the-loop hate speech classification in a multilingual context.

Kunal Relia, Zhengyi Li, Stephanie H Cook, and Rumi Chunara. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 us cities. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, pages 255–264.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.