Machine Learning Based Approach For Hope Speech Detection

S ArunaDevi, B Bharathi

Department of CSE
Sri Sivasubramaniya Nadar College of Engineering,
Tamil Nadu, India

Abstract

In the realm of online communication, hope speech has emerged as a powerful tool capable of reducing hostility and fostering positivity [1]. It offers vital support and inspiration to individuals dealing with illness, stress, loneliness, or depression, helping them feel encouraged and hopeful during challenging times. This paper presents the results of our participation in the "Hope for Equality, Diversity and Inclusion" sub-task of the shared task "HOPE 2024" conducted at IberLEF 2024. The task involved classifying Spanish tweets into 'hope speech' (hs) and 'non-hope speech' (nhs). We employed a logistic regression model, achieving an average macro-F1 score of 0.4161, and ranked 15th in the competition.

Keywords

count vectorizer, TF-IDF, logistic regression, BERT model

1. Introduction

Hope is a unique human ability that allows people to imagine positive future events and outcomes. It helps them stay motivated and optimistic, even in challenging times. These visions significantly affect one's emotions, behaviors, and state of mind. Even if the desired outcome is unlikely, these hopeful visions can have a profound impact on everyday life [2]. In every part of the world, equality, diversity, and inclusion (EDI) have become significant challenges. Language is an essential instrument for communication, and it needs to be inclusive and equitable to all. On social media, however, this is not the case since insulting remarks are made about people based on their nationality, religion, sexual orientation, race, color, or ethnicity. The importance of social media in the lives of the members of vulnerable groups, such as people belonging to the Lesbian, Gay, Bisexual, and Transgender (LGBT) community, racial minorities or people with disabilities, has been studied and it has been found that the social media activities of a vulnerable individual play an essential role in shaping the individual's personality and how he or she views society [3].

Hope plays a crucial role in the well-being, recovery, and restoration of human life [3]. This paper is focused on developing different models that can be used for Hope Speech Detection. For this, we have considered the tweets that promote equality, diversity, and inclusion among the people of the society. From previous studies, we can infer that there exists a 'Snowball Effect' in social media. The Snowball Effect can be given as 'negative comments lead to more negative comments and positive comments lead to more positive comments' [4]. Facebook conducted an experiment by modifying its "Newsfeed" algorithm to show more positive or negative posts to certain users. Their results showed that people tend to write positive posts when they see happy posts in their news feed and vice versa. Therefore, it is important to reinforce positivity on social media by focusing on hope speech [5]. The overview paper for Hope Speech Detection is available in [6]. The overview paper for challenges in natural language processing of Spanish and other Iberian languages is in [7]. The structure of this paper provides a complete understanding of the developing and testing of the models used for Hope Speech Detection.

IberLEF 2024, September 2024, Valladolid, Spain

aruna2210499@ssn.edu.in (S. ArunaDevi); bharathib@ssn.edu.in (B. Bharathi)

This paper is sectioned as follows: Section 2 describes the previous works that have been done by various authors in the field of hope speech detection. Section 3 provides a detailed explanation of the data set. Section 4 provides an overview of the work done.

Section 5 deals with the development of different models for hope speech detection. Section 6 provides details about the evaluation criteria. Section 7 analyses the results obtained from various models. Section 8 provides the conclusion of this paper.

2. Related Works

The Hope Speech Dataset for Equality, Diversity and Inclusion was constructed by Chakravarthi [3], from YouTube comments, in an attempt to encourage research on detecting positive content online. The dataset containing 28,451, 20,198, and 10,705 comments in English, Tamil and Malayalam languages respectively are distributed into two main categories, namely: “Hope” and “Not Hope”. Term Frequency-Inverse Document Frequency (TF-IDF) features were used to train k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Decision Trees (DT), and Logistic Regression (LR) classifiers. Among all the classifiers, DT classifier obtained a weighted-averaged F1-scores of 0.46, 0.51, and 0.56 for English, Tamil, and Malayalam texts respectively.

In their study aimed at identifying hope, Palakodety et al [1] observed that hope exhibited potential in war situations. They provided evidence for this by analyzing multilingual YouTube comments written in both Hindi and English. Their study utilized Logistic Regression with l2 regularization, 80/10 train test split, N-grams (1, 3), sentiment score, and 100-dimensional polyglot FastText embeddings as features, resulting in an F1 score of 78.51. García et al [5] worked on the SpanishHopeEDI dataset which consists of tweets related to LGBT community. They have used different models and concluded that BETO (a BERT model for Spanish) along with Multilayer Perceptron (MLP) performs well in hope speech detection and have achieved an F1 score of 85.12.

Balouchzahi et al [8] proposed a method that utilizes a combination of TF-IDF vectors of words, char sequences, and syntactic n-grams to train: (i) a voting classifier of three estimators, namely: LR, eXtreme Gradient Boosting (XGB), and Multi-Layer Perceptron (MLP) and (ii) Keras Neural Network-based model. They also trained a Bidirectional Encoder Representations from Transformers (BERT) language model from scratch using the given dataset and then used it for Hope Speech detection. For the voting classifier, the authors obtained 1st, 2nd, and 3rd ranks with weighted averaged F1-scores of 0.85, 0.92, and 0.59 for Malayalam, English, and Tamil texts respectively.

Fine-tuning mBERT for Malayalam and Tamil and using BERT for English, Arunima et al [9] obtained weighted-averaged F1-scores of 0.46, 0.81, 0.92 for Tamil, Malayalam, and English texts respectively. Ahani et al [10] emphasize the importance of selecting appropriate algorithms for different languages. They have used KNN and SVM models with TF-IDF features for both English and Spanish dataset. They concluded that SVM performed better on English dataset with an F1 score of 0.49 and KNN performed better on the Spanish dataset having an F1 score of 0.74.

Anusha et al [11] proposed a methodology that addresses the Hope Speech detection by using SMOTE technique to resolve the data imbalance problem and 1D Conv-LSTM model for classification. For English texts, the proposed methodology performed the best and achieved 1st rank with a F1-score of 0.550 but did not perform well for Kannada and Malayalam texts.

3. Dataset Description

The dataset [5][12] for this task was collected between 2020 and 2023. It is an improved and extended version of the SpanishHopeEDI dataset [5]. The version of the dataset for IberLEF 2024 consists of training and development sets on LGTBI-related tweets and a test set on tweets related to the LGTBI collective and other Equality, Diversity, and Inclusion (EDI) topics.

The training dataset is composed of 1,400 tweets in Spanish. There is no class imbalance problem i.e., the number of instances for the first class Hope Speech (hs) is the same as the number of instances for the second class Non-Hope speech (nhs). To conclude, there are 700 instances of data in each of the class. The testing dataset consists of 400 tweets.

4. Proposed work

The proposed methodology consists of two steps. First pre-processing of the data is done to make the data more suitable for the classification model. Pre-processing consists of encoding the categorical variables so that they can be easily used for training the BERT models. Then we used TF-IDF to convert the textual data into numeric format so that the models can classify effectively.

The second step consists of training the models. We have used 4 different machine learning models such as MultinomialNB, SGDClassifier, SVM and Logistic regression. We have also used two different BERT models.

5. Implementation

For the models such as MultinomialNB, SGDClassifier, SVM, and Logistic regression, pre-processing done is the same. For training of the models, we have used K-fold cross-validation technique. In K-fold cross-validation technique, the dataset is split into K number of subsets. Training of the model is done on K-1 subsets while the one subset left is used for the evaluation of the trained model. In this method, we iterate K times over the dataset with a different subset reserved for testing purpose for each iteration. Here we have considered the number of subsets (K) as 10.

Then for each subset of training data and the test data, the TF-IDF vectorizer is used. TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used to convert a collection of raw text documents into a matrix of features. TF (Term Frequency) measures how frequently a term (word) appears in a document. IDF (Inverse Document Frequency) measures the importance of a term across the entire collection of documents. TF-IDF combines both TF and IDF to represent the importance of a term in a specific document relative to its importance across all documents. After finding out the TF-IDF values, we do vectorization. Vectorization involves creating a matrix where each row corresponds to a document and each column corresponds to a term.

For SGDClassifier and logistic regression, the random state is set to 42. For SVM, kernel is to be ’rbf’ (Radial Basis Function) to allow non-linear boundaries. The gamma value is set to 0.5 because a higher gamma value tends to fit the training data very closely, which can lead to overfitting and a lower gamma value generalizes better to unseen data, but may miss fine details in the training data. The C parameter determines the penalty for misclassified data points during the training process. Here C value is set to 1. Now the models are trained on the vectorized training data and accuracy, F-1 score, precision and recall are calculated from the predictions made by the model using the vectorized test data.

Also we have trained 2 Spanish BERT models. The first model that have been used was ’dccuchile/bert-base-Spanish-wwm-cased’ and the other model was ’IIC/bert-base-Spanish-wwm-cased-ctebmsp’. For both models, we have used tokenizer. BERT model uses the tokenizer concept to break some words into sub-words. Also for both the models, training and testing data is classified using train-test-split method from sklearn library with 70% of the dataset belonging to the training dataset and 30% of the dataset belonging to the testing dataset. We have used train-test split instead of K-fold cross-validation because, it is more computationally efficient. Then the training data was tokenized. The categorical

Table 1
Results on training dataset

Model used	Accuracy	F-1 Score	Precision	Recall
MultinomialNB	0.80357	0.808829	0.77866	0.84244
SGDClassifier	0.80142	0.80154	0.79323	0.81133
SVM	0.80571	0.79175	0.84245	0.74798
Logistic Regression	0.80785	0.79727	0.83340	0.76552
BERT 1	0.81472	0.79896	0.89595	0.72093
BERT 2	0.80337	0.83253	0.85714	0.80930

variables of the train data 'hs' and 'nhs' have been encoded to 0 and 1 respectively.

A data loader is used for the training data. Dataloader from Pytorch is used to divide the data into batches and also to shuffle the data. Batching allows the model to process data in smaller chunks (batches) instead of the entire dataset at once. Here data is divided into batches of size 30. Shuffling the data order in each epoch ensures the model encounters data points in different combinations, forcing it to learn generalizable features rather than memorizing specific data orders. Optimizer is used to minimize the model's error thereby enhancing the model's performance. The data is trained in the model for 2 epochs. The model predicts the output of the training data and loss is calculated and propagated backward. The implementation details are available in GitHub [1].

Then the testing data is fed into the models after tokenization. The predicted output from the models are mapped into their respective categories and stored in the file. Then accuracy, F-1 score, precision score, and recall score are calculated. The accuracy, F-1 score, precision score, and recall score of all the models that have been described previously are tabulated in Table 1.

6. Evaluation

Precision is defined as the ratio of correctly classified positive samples to a total number of classified positive samples (either correctly or incorrectly).

Precision = TP / (TP+FP)
where,
TP = number of samples correctly predicted as positive
FP= number of samples incorrectly predicted as positive

The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

Recall = TP / (TP+FN)
where,
TP = number of samples correctly predicted as positive
FN= number of samples incorrectly predicted as negative

The F1 score is calculated as the harmonic mean of precision and recall. The value of the F1 score lies between 0 to 1.

F1 score= (2*precision*recall) / (precision+recall)

[1] Implementation details repository: https://github.com/S-ArunaDevi06/hope_speech_detection_Spanish

Table 2
Results on testing dataset

Model used	Accuracy	F-1 Score	Precision	Recall
BERT 1	0.55	0.24561	0.75	0.15
MultinomialNB	0.545	0.50106	0.5542	0.46
Logistic Regression	0.5725	0.4161	0.665	0.305

The macro F1 score calculates the unweighted mean of the F1 scores calculated for each class. This task's evaluation measure is a macro F1 score.

7. Result Analysis

From Table 1, we can conclude that the accuracy and the F-1 scores of BERT 1, MultinomialNB, and Logistic regression are quite high. These models were submitted. And on the testing data, Logistic regression performed well with an accuracy of 0.5725. So it was made as the final submission. It had an average Macro F-1 score of 0.4161. Using the test data with the correct labels, which was provided after the competition, precision, recall, accuracy, and average macro F1 scores of all the 3 models which are mentioned earlier are calculated and tabulated in Table 2. From this, we can conclude that Logistic regression has higher accuracy and lower F1 score than MultinomialNB.

By looking at the misclassified sentences from the predictions made by Logistic regression, we can infer that some sentences such as "Hace historia como la primer mujer trans ganadora de un #GoldenGlobe y hay que adorarla" are misclassified as the model fails to capture complex emotions such as admiration and pride. And sentences involving idiomatic expressions such as "Salir del closet es un acto de valentía" are often misclassified.

8. Conclusion

In the era of ubiquitous internet, public opinion on a rapidly evolving global issue can exhibit similar fast-changing behavior, much of which is visible to a very large fraction of internet users. In this work, we have developed models for hope speech detection. Our approach consisted of two processes: pre-processing of the data and training of the model. We have used 4 traditional machine learning models and also trained 2 BERT models. The best-performing model on the testing data was Logistic regression with an accuracy of 0.5725 and an average Macro F-1 score of 0.4161 had ranked 15th in the task.

References

[1] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, arXiv preprint arXiv:1909.12940 (2019).

[2] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078. URL: https://www.sciencedirect.com/science/article/pii/S0957417423005808. doi:https://doi.org/10.1016/j.eswa.2023.120078.

[3] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: M. Nissim, V. Patti, B. Plank, E. Durmus (Eds.), Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[4] A. Sundar, A. Ramakrishnan, A. Balaji, T. Durairaj, Hope speech detection for dravidian languages

using cross-lingual embeddings with stacked encoder architecture, SN Computer Science 3 (2022) 67.

[5] D. García-Baena, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgbt case, Language Resources and Evaluation 57 (2023) 1487–1514.

[6] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. Lambebo Tonja, J. A. García-Díaz, S. Bozkurt, B. R. Chakravarthi, H. G. Ceballos, V.-G. Rafael, G. Sidorov, L. A. Ureña-López, A. Gelbukh, S. M. Jiménez-Zafra, Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations, Procesamiento del Lenguaje Natural 73 (2024).

[7] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[8] F. Balouchzahi, B. Aparna, H. Shashirekha, MUCS@ LT-EDI-EACL2021: Cohope-hope speech detection for equality, diversity, and inclusion in code-mixed texts, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 180–187.

[9] S. Arunima, A. Ramakrishnan, A. Balaji, D. Thenmozhi, et al., ssn_dibertsity@ LT-EDI-EACL2021: hope speech detection on multilingual youtube comments via transformer based approach, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 92–97.

[10] Z. Ahani, G. Sidorov, O. Kolesnikova, A. Gelbukh, Zavira at HOPE2023@ IberLEF: Hope speech detection from text using TF-IDF features and machine learning algorithms (2023).

[11] A. Gowda, F. Balouchzahi, H. Shashirekha, G. Sidorov, MUCIC@ LT-EDI-ACL2022: Hope speech detection using data re-sampling and 1D Conv-LSTM, in: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp. 161–166.

[12] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078. doi:10.1016/j.eswa.2023.120078.