

NLP_SSN_CSE at HOPE2023@IberLEF : Multilingual Hope Speech Detection using Machine Learning Algorithms

Varsha Balaji, Aishwarya Kannan , Aishwarya Balaji and Bharathi Bhagavath Singh

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

Abstract

Lately, there has been an issue of vulgarity and negative comments on social media platforms like YouTube, Twitter, and Instagram. Offensive comments lead to conflicts among the users. This in turn hinders the reach of the positive aspects of social media to the people. The given task was to classify the data as hope or non-hope speech. YouTube comments and tweets that provided hope, positivity, and equality and those that did not provide these were used for the English and Spanish dataset respectively. To classify the data we used several machine learning models such as BERT:bert-base-multilingual-uncased and bert-base-uncased, Random Forest, SVM, Logistic Regression, and Decision Tree. Out of these, mBERT produced the best results.

1. Introduction

In today's world people face many challenges which include racism, sexism, and other religion-related issues such as caste. The present generation finds it easier to come out on online platforms such as YouTube, Instagram, and Twitter, looking for hope from strangers. The comments of these strangers have a profound impact on the individual. The comment can either motivate the individual or pull him down. So it is important to make sure it has a positive impact on the user.

YouTube is a diverse platform where people across the globe can create and post videos. The user can make a video on any topic and may monetize the video. The viewers can like, share and comment on the video. The content of this media increases the knowledge of the viewers and makes them aware of the current happenings. Twitter is also a social media platform used to connect with people. It can be used to receive news or follow the people we like.

Recently there has been a lot of research going on related to Hope Speech Detection. Hope speech depends on motivation and hope from social media platforms[1]. So it is important to generate a solution that helps in promoting hope to the people by showing positive tweets and

IberLEF 2023, September 2023, Jaén, Spain


✉ varsha2010399@ssn.edu.in (V. Balaji); aishwarya2010864@ssn.edu.in (A. K.); aishwarya2010142@ssn.edu.in (A. B.); bharathib@ssn.edu.in (B. B. S.)

🌐 <https://www.ssn.edu.in/staff-members/dr-b-bharathi/> (B. B. S.)

🆔 0000-0001-7279-5357 (B. B. S.)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

comments while desisting the negative ones.

In this paper a part of [2], we have addressed Hope Speech Detection using some of the pre-existing transformer models.[3] We have obtained the dataset from the tweets and YouTube comments in English, and Spanish. We have used basic models like Logistic Regression, SVM, Decision Tree, Random Forest, and multilingual transformers such as BERT. We have achieved a good result using the multilingual BERT transformer model. Hence the task was solved using it. The overview of the HOPE2023@IberLEF: Multilingual Hope Speech Detection task is given in [2].

The remaining part of this paper is organized as follows. Section 2 discusses the other related works on the Hope Speech Detection task. The dataset for this task is discussed in Section 3. Sections 4 and 5 touch upon the features and methods used for this task. Results are written in Section 6. Section 7 conveys the conclusion of the paper.

2. Related Work

A lot of research work has been carried out to deal with Hope Speech Detection. The paper Poly-Hope: Two-level hope speech detection from tweets [4], [5],[6] focuses on the classification of the speech(binary and multi class). Eight traditional machine learning classifiers were used, namely: LR, SVM with Radial Basis Function (RBF), and linear kernels, RFC, XGB, AdaBoost, and Catboost, are used for the hope speech detection task. All the classifiers were used with default parameters and were trained on the TF-IDF vectors of word uni-grams. Sentence transformer-based hope speech detection for Equality, Diversity, and Inclusion is described in [7]. Hope Speech Detection on multilingual YouTube comments via transformer based approach: a paper on hope speech classification [8]. The classification was done for three languages: English, Tamil, and Malayalam. Traditional models like SVM, Logistic Regressions, and Naive Bayes as well as transformers like MT5 and BERT were used. Promising results were obtained using multilingual BERT for Tamil and Malayalam, and BERT for English YouTube comments. Hope Speech Detection using Machine Learning [9]: Here the balanced data was first passed through machine learning classifiers. Further, deep learning techniques like DNN, DNN with embedding (DNN+Emb), CNN, LSTM, and BiLSTM were applied. The RF model achieved the best performance with over-sampled data.

3. Dataset Analysis and Preprocessing

Emojis, abbreviations, and small words are all permitted in YouTube comments. The data must first be processed before being trained. Any machine learning solution must use a Language Training Development Test including data pre-processing to be successful. Many YouTube comments may contain misspelled words and indications of inconsistent text continuity. Pre-processing is the removal of all HTML tags, hashtags, social media mentions, and URLs in order to clean up the dataset and normalize these abnormalities. Emojis and emoticons, which are crucial in characterizing the speech, must also be annotated. These are

| Language | Training | Development | Testing |
|----------|----------|-------------|---------|
| English | 22651 | 2799 | 4805 |
| Spanish | 1312 | 300 | 450 |

Table 1
Dataset Description for English and Spanish

| Model | Label |
|---|-------|
| Apartándonos un poco del hype, hoy celebramos y nos {unimos al #OrgulloLGTBI también desde DLH | HS |
| Naaa pero son los putos fachas fascistas que quieren quitar derechos y {libertad de expresión, #SpainIsAFascistState #LGTB Republica | NHS |

Table 2
Dataset sample for Spanish

| Model | Label |
|--|-------|
| these tiktoks radiate gay chaotic energy and I love it | NHS |
| Network Engineer here- 23 and currently working as an instructor teaching {men and women looking to be in IT =} Next I want to teach at a University! | HS |

Table 3
Dataset sample for English

taken out of the comment and replaced with the text they stand for. Short terms that may be present in the text data are replaced with the full version of such words. We use a look-up table to change short words into their extended forms, such as "what's" becoming "what is" and "u" becoming "you." After that, the series of texts are changed to lowercase, and any extraneous white spaces are eliminated.

[10] The Natural Language Toolkit (NLTK) was used in data preprocessing of the hope speech tasks in natural language processing (NLP). For the study of hope speech, it provides a variety of techniques and functions, including tokenization, stemming, lemmatization, and part-of-speech tagging. With the use of NLTK, unstructured text data may be converted into a format that is appropriate for NLP modeling and analysis, giving researchers and developers the ability to efficiently handle and analyze data from hope speech. The creation of speech-related applications and NLP research benefit greatly from NLTK's extensive library of corpora, lexical resources, and algorithms.

3.1. Acquired Dataset

Table 1 shows training, development, and testing data for a multilingual hope speech detection system using machine learning. It includes statistics for English (trained on 22,651, developed on 2,799, tested on 4,805) and Spanish (trained on 1,312, developed on 300, tested on 450).

The tables 2 & 3 below presents a description of the dataset examples, encompassing both hope and non-hope speech instances in English and Spanish[11]. The table 2 has two

instances with the label "Model Label." The first instance is an emotion that is supportive of homosexual wild TikTok films and expresses affection for them. The second example is a resume for a network engineer who works as a teacher and expresses a desire to teach at a university. The two instances have nothing to do with the goal of the hope speech.

Additionally, the table 3 includes two samples with the heading "Model Label." In the first illustration, people get together to celebrate and show their support for LGBTQ+ people. The second illustration emphasizes opposition to fascist beliefs and support for free speech. These instances emphasise inclusion and democratic principles while addressing the issue of hope. The aim was to distinguish between the hopeful and uninspired.

4. Feature Extraction

The process of turning raw data into a collection of pertinent characteristics that can be utilized as input for machine learning models is known as feature extraction. This is frequently applied in disciplines like image identification, natural language processing, and others where numerical visualization of data is required.

4.1. Count Vectorizer

Text data is transformed into numerical feature vectors using the feature extraction approach known as CountVectorizer in natural language processing. It operates by calculating a sparse matrix from the frequency counts of each word in a document. Several machine-learning models can be entered into the resulting matrix. The representation of text data as a numerical input for machine learning models is simple yet efficient. [12] uses CountVectorizer in fake news detection task that helps in improving the final accuracy of the model in which CountVectorizer has been applied to.

4.2. TF-IDF Vectorizer

Term frequency-inverse document frequency, or TF-IDF The feature extraction method known as "Vectorizer" is frequently applied in natural language processing to transform text data into numerical feature vectors. It operates by evaluating each word's significance in relation to the corpus of documents as a whole. The resulting matrix gives each word's frequency and weight in each document in numerical form. The TF-IDF Vectorizer can be used as input for various machine learning models and is helpful for locating significant words in a document. The settings halt word removal and minimum document frequency can both be changed.

4.3. BERT Encoding

The pre-trained BERT model was adjusted on the provided dataset of hope speech in order to apply BERT (Bidirectional Encoder Representations from Transformers) encoding for Hope-speech classification. By feeding the dataset to the BERT model, which creates contextualized

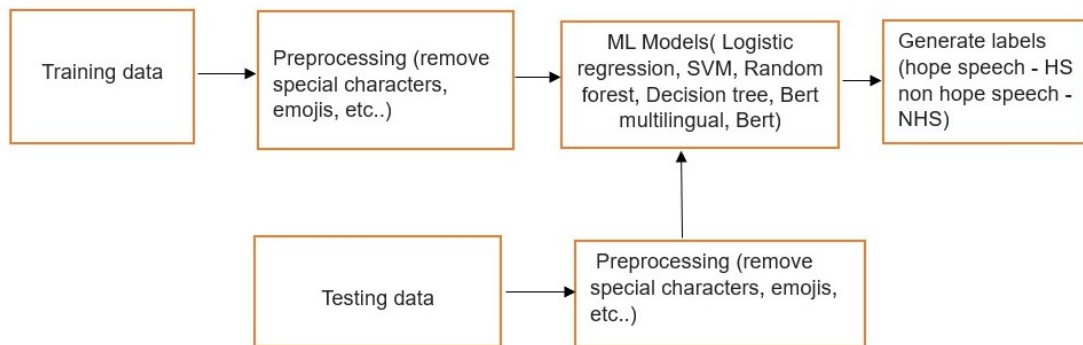


Figure 1: Methodology

word embeddings for each input sentence, this can be accomplished. Following that, a classification model—such as a neural network—learns to predict whether a given text contains hate speech or not using the generated embeddings as input. When compared to simpler feature extraction techniques, BERT’s capacity to include contextualized meaning can help increase the accuracy of hope-speech classification. A model for detecting hate speech and hope speech in text data can be created by fine-tuning BERT for a particular job of hope speech classification.

5. Methodology

5.1. Random Forest classifier

A Random Forest classifier is a meta-estimator that employs averaging to increase predicted accuracy and reduce over-fitting after fitting numerous decision tree classifiers to distinct dataset subsamples. Each decision tree in the Random forest model is built using a subset of characteristics and a subset of data points. Simply described, the data set containing k records is divided into n random records and m features. For the samples, individual decision trees are constructed generating the specific outputs. The resultant output of the data set samples is generated based on averaging. The model trained for the given dataset generated an accuracy of 92.40% with 90% F1-score and 91% precision.

5.2. SVM

SVM (Support Vector Machines) algorithms can be used for both regression and classification problems. The given dataset is a classification-based problem, A model is created by an SVM classifier that classes fresh data points into one of the predetermined categories. As a result, it can be thought of as a binary linear non-probabilistic classifier. SVMs are applicable to linear classification tasks. The model trained for the given dataset generated an accuracy of 89.12%

with 90% F1-score and 89% precision for the English language. Using the kernel approach, SVMs may effectively do non-linear classification in addition to linear classification. It allows us to automatically map the inputs into large feature areas.

5.3. Logistic Regression

Machine learning uses the categorization method known as Logistic regression. The dependent variable is modeled using a logistic function. Because the dependent variable is dichotomous, there are only two conceivable classifications it could belong to (for example, cancer can either be malignant or not). This method is therefore employed while working with binary data. The sigmoid function is used in logistic regression to convert predicted values to probabilities. This function turns any real value between 0 and 1 into another value. The model trained for the given dataset generated an accuracy of 92.10% with 90% F1-score and 91% precision for the English language.

5.4. Decision Tree

The non-parametric supervised learning approach used for classification and regression applications is the Decision Tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes. By using a greedy search to find the ideal split points inside a tree, decision tree learning uses a divide-and-conquer technique. When most or all of the records have been classified under distinct class labels, this splitting procedure is then repeated in a top-down, recursive fashion. The model trained for the given dataset generated an accuracy of 89.91% with 90% F1-score and 90% precision for the English language.

5.5. BERT : bert-base-multilingual-uncased and bert-base-uncased

It is a pre-trained model that was first described in the publication BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [13]. It was trained using a masked language modeling (MLM) goal on the top 102 languages with the largest Wikipedia. It is a transformer model that has been previously self-supervised by trained on a sizable corpus of multilingual data. BERT employs bi-directional learning to simultaneously understand word context from the left to the right. The [14] Masked Language Modelling (MLM) technique, which involves randomly masking 15% of the input's words before putting it through the model to forecast the masked words, is best suited for this bidirectional approach. Additionally, it aids in the optimization of Next Sentence Prediction (NSP), which foretells the relationship (whether they will follow one another or not) between two phrases. The bert-base-uncased model trained for the given dataset generated an accuracy of 89.91% with 90% F1-score and 90% precision for the English language.

6. Observation

In this part, we will examine how well different machine learning supervised models perform for the two languages (English and Spanish). The weighted F1 score determines the excellence

| Model | Feature Extraction | Precision | Recall | F1 score | Accuracy |
|--------------------------------|--------------------|-----------|--------|----------|----------|
| Random Forest | Count Vectorizer | 91 | 92 | 90 | 92.40 |
| Decision Tree | Count Vectorizer | 89 | 89 | 89 | 89.21 |
| Logistic Regression | Count Vectorizer | 90 | 92 | 91 | 91.80 |
| SVM | Count Vectorizer | 89 | 90 | 90 | 89.12 |
| Random Forest | TFIDF Vectorizer | 91 | 92 | 89 | 92.07 |
| Decision Tree | TFIDF Vectorizer | 90 | 90 | 90 | 89.91 |
| Logistic Regression | TFIDF Vectorizer | 91 | 92 | 90 | 92.10 |
| SVM | TFIDF Vectorizer | 90 | 92 | 90 | 91.94 |
| bert-base-multilingual-uncased | Count Vectorizer | 92 | 93 | 91 | 92.87 |
| bert-base-uncased | Count Vectorizer | 83 | 87 | 91 | 91.26 |

Table 4
Classification report for English Train dataset with 80-20 Train test split

| Model | Feature Extraction | Precision | Recall | F1 score | Accuracy |
|--------------------------------|--------------------|-----------|--------|----------|----------|
| Random Forest | Count Vectorizer | 76 | 76 | 76 | 75.66 |
| Decision Tree | Count Vectorizer | 67 | 67 | 67 | 66.53 |
| Logistic Regression | Count Vectorizer | 75 | 75 | 75 | 75.20 |
| SVM | Count Vectorizer | 73 | 73 | 73 | 73.38 |
| Random Forest | TFIDF Vectorizer | 83 | 91 | 87 | 91.26 |
| Decision Tree | TFIDF Vectorizer | 92 | 93 | 91 | 92.87 |
| Logistic Regression | TFIDF Vectorizer | 77 | 77 | 77 | 77.18 |
| SVM | TFIDF Vectorizer | 77 | 77 | 77 | 76.80 |
| bert-base-multilingual-uncased | Count Vectorizer | 97 | 97 | 97 | 96.57 |
| bert-base-uncased | Count Vectorizer | 86 | 86 | 85 | 85.55 |

Table 5
Classification report for Spanish Train dataset with 80-20 Train test split

of the models. The tables below present the evaluation results of all the models on the training dataset. The model used to predict accuracy for the training dataset include Random Forest classifier, SVM, Logistic Regression, and Decision Tree with respect to the classification algorithms, and transformer BERT models including the bert-base-multilingual-uncased and bert-base-uncased. Among all the models trained, bert-base-multilingual-uncased gave the optimal results for the English and Spanish dataset with a weighted F1 score of 92.87% and 96.57% respectively. The Logistic Regression and Random Forest classifiers have provided similar F1 scores of 92.10% and 92.07% for the English dataset using the TF-IDF vectorizer. However, the results obtained for Spanish were comparatively low.

The tables given above depict the classification report of various classification models that were obtained for the training dataset. Tables 4 and 5 represent the results and accuracy obtained for the training dataset. Among all the models tested for the training dataset BERT model, resulting in better. The tables below present the evaluation results of all the models on the test dataset.

The test dataset as in table 6 results generated an F1 score of 0.5913 for Spanish and 0.4937 for English. This model can be further improved to deal with data in multiple languages in the future.

| Model | Language | Feature Extraction | F1 score |
|--------------------------------|----------|--------------------|----------|
| bert-base-multilingual-uncased | English | Count Vectorizer | 0.4937 |
| bert-base-multilingual-uncased | Spanish | Count Vectorizer | 0.5913 |

Table 6
Test dataset results

We secured eighth and fifth positions in the leader board for the Spanish and English dataset respectively. The F1 score for the Spanish dataset was 59.14% wherein the highest was 91.61% and the F1 score for the English dataset was 49.37% wherein the highest was 50.12%.

7. Conclusion

The requirement for Hope Speech Detection in social media is increasing. Nearly 75% of connections today happen to be online. This seeks to differentiate between hope and non-hope speech to promote a good atmosphere and shape human minds instead of making them feel bad and low about themselves. Hope Speech Detection models, although important, have inadequate amounts of research done on them. In this paper, pre-trained multilingual transformer models are used to detect Hope Speech in 2 languages, namely English and Spanish.

References

- [1] B. R. Chakravarthi, Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, 2020, pp. 41–53.
- [2] S. M. Jiménez-Zafra, M. Á. García-Cumbreras, D. García-Baena, J. A. García-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection, *Procesamiento del Lenguaje Natural* 71 (2023).
- [3] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.
- [4] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, *Expert Systems with Applications* (2023) 120078.
- [5] C. R. Snyder, S. J. Lopez, H. S. Shorey, K. L. Rand, D. B. Feldman, Hope theory, measurements, and applications to school psychology., *School psychology quarterly* 18 (2003) 122.
- [6] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, et al., Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in:

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 378–388.

- [7] B. Bharathi, D. Srinivasan, J. Varsha, T. Durairaj, B. Senthilkumar, Ssnscse_nlp@lt-edi-acl2022: hope speech detection for equality, diversity and inclusion using sentence transformers, in: LTEDI, 2022.
- [8] S. Arunima, A. Ramakrishnan, A. Balaji, D. Thenmozhi, et al., ssn_dibersity@lt-edi-acl2021: 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.
- [9] P. Roy, S. Bhawal, A. Kumar, B. R. Chakravarthi, Iiitsurat@lt-edi-acl2022: Hope speech detection using machine learning, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 120–126.
- [10] E. Loper, S. Bird, Nltk: The natural language toolkit, arXiv preprint cs/0205028 (2002).
- [11] 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 (2023) 1–28.
- [12] A. Patel, K. Meehan, Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinomialnb and support vector machine, in: 2021 32nd Irish Signals and Systems Conference (ISSC), IEEE, 2021, pp. 1–6.
- [13] F. Souza, R. Nogueira, R. Lotufo, Bertimbau: pretrained bert models for brazilian portuguese, in: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9, Springer, 2020, pp. 403–417.
- [14] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint arXiv:1906.01502 (2019).