Zavira at HOPE2023@IberLEF: Hope Speech Detection from Text using TF-IDF Features and Machine Learning Algorithms

Zahra Ahani, Grigori Sidorov , Olga Kolesnikova and Alexander Gelbukh

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico city, Mexico

Abstract

This paper presents the results of our participation in the shared task Multilingual Hope Speech detection aimed at classifying texts into hope and non-hope categories. The task involved two datasets, one in English and the other in Spanish. We used the SVM algorithm for the English data and the KNN algorithm for the Spanish data. Our approach achieved the third place on both datasets. Specifically, our SVM-based approach achieved an F1 score of 0.49, while our KNN-based approach achieved an F1 score of 0.74. Our results suggest that cross-lingual classification of hope and non-hope texts is a challenging task, particularly due to the linguistic differences between languages. Nevertheless, our results demonstrate the effectiveness of the SVM and KNN algorithms for this task, highlighting the importance of selecting appropriate algorithms for different languages. Overall, this paper contributes to the growing body of research on cross-lingual text classification and provides insights for future work in this area.

Keywords

Hope Speech, TF-IDF, Machine Learning

1. Introduction

Hope is a unique ability possessed by humans that allows them to imagine possible future scenarios and their potential outcomes with adaptability [1]. Such imagination can have a significant impact on a person's behaviors [2]. The aim of hope speech is to communicate the conviction that an individual can be inspired to progress in life and attain her aspirations [3].

Online social media platforms have a considerable impact on human existence, with individuals expressing their opinions openly. The notable characteristics of Social Media, including swift distribution, affordability, availability, and anonymity, have contributed to the popularity of these platforms. As social media offers data to develop profound understanding of human behavior on these platforms, they have become significant sources for research on Natural Language Processing (NLP) issues [4, 5, 6].

With the recent technological advancements in social media, people's daily routines have expanded to include virtual interactions and networks. As a result, social media platforms have a

IberLEF 2023, September 2023, Jaén, Spain

zahraahani97100@gmail.com (Z. Ahani); sidorov@cic.ipn.mx (G. S.); kolesnikova@cic.ipn.mx (O. Kolesnikova); gelbukh@cic.ipn.mx (A. Gelbukh)

https://www.researchgate.net/profile/Olga-Kolesnikova-2 (O. Kolesnikova)

0000-0002-1307-1647 (O. Kolesnikova)

significant impact on users' daily lives [7]. Many users share positive and hopeful messages with the aim of inspiring peace [8]. Numerous internet forums have gained popularity as a means of offering assistance, guidance, and emotional support. Additionally, when users experience challenging or unfavorable circumstances, they may turn to virtual platforms, in addition to seeking support from their loved ones, for help and guidance [9, 10].

Currently several tasks such as sentiment analysis [11], hate speech, [4] and fake news [7] are well-explored text classification tasks and compared to them, hope speech detection is relevantly under explored [12]. The identification of hope speech involves analyzing and detecting optimistic discussions, comments, and posts that convey positive sentiments, such as promoting adherence to COVID-19 guidelines, for example. Hope speech is distinguished from messages that express discriminatory attitudes towards groups with non-heterosexual orientations, such as Lesbian, Gay, and Transgender individuals [13].

The task is defined as a binary classification challenge, with the objective of identifying instances belonging to either the "Hope" or "Non-Hope" categories [14, 15, 3]. To address this task, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) were employed as machine learning approaches, and TF-IDF vectorization was utilized as a feature engineering technique.

SVM is a supervised machine learning technique that is effective for both data classification and regression [16], as it has showed in other natural language tasks [17].The SVM objective is to identify the hyperplane in an N-dimensional space that can accurately separate the data points. This means that the algorithm can draw a decision boundary line that distinguishes between data points belonging to one category and data points of the other category. This technique is applicable to almost all vector-encoded data, provided such encoding is efficient. By creating a good vector representation of our data, we can obtain satisfactory results using SVM. On the other hand, KNN (K-Nearest Neighbors) is a non-parametric supervised machine learning algorithm that classifies a new data point based on the majority class of its k-nearest neighbors in the training data. [18] To represent the texts in vector form, we used TF-IDF (Term Frequency - Inverse Document Frequency) vectorization. it is a feature engineering that helps transforming text data into numerical vectors. it assigns weigth to each term in a document based on how important it is to the overall meaning of that document. [19]

2. Literature Review

Chakravarthi [3] pioneered the field of hope speech detection on social media platforms by creating the HopeEDI corpus using YouTube comments in both Dravidian and English languages. Initially the corpus included English as well as code-mixed Tamil-English and Malayalam-English datasets, and it was later expanded to include Spanish and code-mixed Kannada-English texts Chakravarthi et al. [20]

In their study aimed at identifying hope, Palakodety et al. [21] observed that hope exhibited potential in war situations. They provided evidence for this by analyzing multilingual YouTube comments written in both Hindi and English, using Devanagari and Roman script. 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 F-1 score of 78.51 (2.24%).

Balouchzahi et al. [8] proposed a method for Hope Speech detection that utilizes a combination of TF-IDF vectors of words, char sequences, and syntactic n-grams to train a voting classifier and a Keras Neural Network-based model. They also trained a BERT language model from scratch and obtained high F1-scores for Malayalam and English texts, but a lower score for Tamil texts.

Dowlagar and Mamidi [21] used multilingual BERT embedding for CNN classifier after preprocessing texts, obtaining high rankings in Hope Speech detection for Tamil, Malayalam, and English texts. Arunima et al. [22] fine-tuned mBERT for Malayalam and Tamil and used BERT for English to obtain high weighted-averaged F1-scores for Tamil, Malayalam, and English texts. Upadhyay et al. [23] tried two approaches, using contextual embeddings with classifiers and a majority voting ensemble of BERT, RoBERTa, ALBERT, and LSTM models to achieve high weighted-averaged F1-scores for English, Malayalam, and Tamil texts.

To address the challenge of language identification in code-mixed data, Shahiki Tash et al. [24] used SVM and KNN, as well as the TF-IDF vectorizer for feature extraction. they also highlighted the importance of language identification in code-mixed text and the use of machine learning techniques for that purpose.

3. Dataset

The dataset consists of two collections of data, one in Spanish and the other in English, which were collected from 2019 to 2022 [15]. The Spanish collection is a larger version of the Spanish HopeEDI dataset and was used in the ACL LT-EDI-2022 Spanish task [3], while the English collection is a part of the HopeEDI dataset. Both collections contain tweets and YouTube comments on various social topics, and the comments are labeled as either Hope Speech (HS) or Non-Hope Speech (NHS). In the Spanish corpus, tweets that promote social integration, inspire the LGTBI community, encourage LGTBI people, or advocate for tolerance are labeled as HS, while tweets that express negative sentiment, promote violence, or use gender-based insults are labeled as NHS. Chakravarthi et al. [20] The dataset comprises approximately 2,550 Spanish tweets and 28,424 English YouTube comments (see, Table1)[25, 26].

Language	Split	Hope	Non Hope	Total
English	Train data	2229	23221	25450
	Test data	21	4784	4805
Spanish	Train data	791	821	1612
	Test data	150	300	450

Table 1
Train and test data in English and Spanish language

4. Methodology

In this study, we explored the effectiveness of two machine learning algorithms, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), on text classification tasks in English and

Spanish. Our approach involved several stages of data preprocessing and feature engineering in order to optimize the performance of the models. To evaluate the performance of our binary classification models, F1-score was used. The F1 score is a metric that combines precision and recall to assess the performance of a classification model. It provides a balanced measure of accuracy by taking into account both the model's ability to make accurate positive predictions (precision) and its ability to correctly identify positive instances (recall). The F1 score ranges from 0 to 1, with higher values indicating better performance. It is a useful tool for comparing models and evaluating the overall effectiveness of classification algorithms. This measure combines the precision and recall of a classifier into a single metric by taking their harmonic mean.

It is defined as:

$$F1 = \frac{2(Precision \times recall)}{(Precision + Recall)} \tag{1}$$

4.1. Pre-processing

The first stage of our methodology involved extensive pre-processing of the textual data. This included the use of a lemmatizer to reduce words to their base form and the removal of stop words and punctuation marks to simplify the text. Additionally, we employed a clean-text method to remove any irrelevant or redundant information noise that might negatively impact the performance of the models.

4.2. Feature Engineering

To extract features, we employed the Scikit-learn module's TF-IDF Vectorizer to extract character n-grams from pre-processed textual data that was previously cleaned and lemmatized. The TF-IDF formula is a commonly used weighting scheme in information retrieval and text mining. It quantifies the importance of a term within a document relative to a collection of documents. The formula is composed of two components: the term frequency (TF) and the inverse document frequency (IDF). Term frequency (TF) measures the occurrence of a term within a document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in that document. Table 2 provides a list of the parameters we used for the TF-IDF process. The vectorized data was split into training and testing subsets using Scikit-learn's train_test_split function.[27]

TF-IDF Formula:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) \tag{2}$$

where:

$$\text{tf}(t, d) = \frac{\text{number of times term } t \text{ appears in document } d}{\text{total number of terms in document } d}$$

$$\text{idf}(t) = \log\left(\frac{\text{total number of documents}}{\text{number of documents containing term } t}\right)$$

4.3. Model Construction

Once the TF-IDF vectorization was applied, each document was transformed into a vector within a high-dimensional space. In this space, the dimensions represented the distinct terms found in the corpus. In addition, we varied certain parameters in the models and vectorized. Table 2 corresponds to the best-performing approaches. The reason for selecting these specific hyperparameters is that they have been found to yield better results compared to other parameter choices in our experiments. Through iterative experimentation and evaluation, we observed that these particular parameter settings led to improved performance, we measured through metrics such as accuracy and F1 score, on this dataset .

Name of classifier\vectorizer	Parameter1	Parameter2	Parameter3	Parameter4
SVM (English)	C=1	kernel='poly'	degree=2	gamma='scale'
KNN (Spain)	n_neighbors=6	metric='minkowski'	p=2	weights='uniform'
TF-IDF	min_df=0	ngram_range=(2,3)	analyzer='char'	input='content'

Table 2
Hyper parameters used in the experiments

5. Results

In this study, we employed SVM and KNN algorithms to analyze data in English and Spanish languages. The results, shown in Table 3, revealed that the KNN model performed exceptionally well, achieving an F1 HS score of 0.67 for the Spanish data. However, we did not obtain satisfactory results with the SVM algorithm for this dataset. It is worth noting that the performance of the Spanish data was better than the English data, indicating a well-balanced dataset.

For the English dataset, we also applied SVM and KNN algorithms. Interestingly, we obtained better results with the SVM algorithm compared to the KNN algorithm. However, none of the participants in the task achieved a prediction higher than 1 percent for F1 HS, suggesting that the data was not adequately balanced.

Specifically, the SVM algorithm yielded an F1 score of 0.4975 for the English data, while the KNN algorithm resulted in an F1 score of 0.7430 for the Spanish data. These findings highlight the effectiveness of the machine learning techniques employed in the study. Furthermore, they underscore the importance of balancing the data properly to obtain accurate and reliable results.

Participants	Language	Average Macro F1	Precision HS	Recall HS	F1 HS	Precision NHS	Recall NHS	F1 NHS
1	Spanish	0.9161	0.8671	0.9133	0.8896	0.9555	0.9300	0.9426
2	Spanish	0.7437	0.9091	0.4667	0.6167	0.7855	0.9767	0.8707
My result	Spanish	0.7430	0.6215	0.7333	0.6728	0.8535	0.7767	0.8133
4	Spanish	0.7238	0.5864	0.7467	0.6569	0.8533	0.7367	0.7907
1	English	0.5012	0.0163	0.1905	0.0301	0.9963	0.9496	0.9724
2	English	0.4989	0.0000	0.0000	0.0000	0.9956	1.0000	0.9978
My result	English	0.4975	0.0000	0.0000	0.0000	0.9956	0.9944	0.9950
4	English	0.4974	0.0000	0.0000	0.0000	0.9956	0.9941	0.9949

Table 3
Top 4 results of share task

6. Conclusion

Based on our methodology of using KNN and SVM models, and classical NLP strategy with TF-IDF features, we were able to effectively classify textual data in both English and Spanish.

Our SVM-based approach achieved an F1 score of 0.49 in English data, while our KNN-based approach achieved an F1 score of 0.74 in Spanish data, this emphasizes the importance of selecting appropriate algorithms for different languages. Also extensive experimentation was conducted on the given dataset, resulting in varied outcomes. A significant finding that contributed to performance improvement was the meticulous selection of optimal hyperparameters through iterative testing. By repeating the experiments with different hyperparameter configurations, it became evident that choosing appropriate hyperparameters consistently led to enhanced results. Our results also highlight the challenges of the cross-lingual classification of hope and non-hope texts due to linguistic differences between languages. Our approach contributes to the growing body of research on cross-lingual text classification and provides valuable insights for future work in this area.

7. Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications (2023) 120078.

[2] P. Bruininks, B. F. Malle, Distinguishing hope from optimism and related affective states, Motivation and emotion 29 (2005) 324–352.

[3] 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.

[4] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, Hssd: Hate speech spreader detection using n-grams and voting classifier, in: CEUR Workshop Proceedings, volume 2936, CEUR-WS, 2021, pp. 1829–1836.

[5] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbuk, Transformer-based model for word level language identification in code-mixed kannada-english texts, arXiv preprint arXiv:2211.14459 (2022).

[6] M. Gemeda Yigezu, A. Lambebo Tonja, O. Kolesnikova, M. Shahiki Tash, G. Sidorov, A. Gelbukh, Word level language identification in code-mixed Kannada-English texts

using deep learning approach, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 29–33. URL: https://aclanthology.org/2022.icon-wlli.6.

[7] F. Balouchzahi, H. Shashirekha, G. Sidorov, Mucic at checkthat! 2021: Fado-fake news detection and domain identification using transformers ensembling, in: 2021 Working Notes of CLEF-Conference and Labs of the Evaluation Forum, CLEF-WN 2021, 2021, pp. 455–464.

[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] N. Ghanghor, P. Krishnamurthy, S. Thavareesan, R. Priyadharshini, B. R. Chakravarthi, Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada, in: Proceedings of the first workshop on speech and language technologies for dravidian languages, 2021, pp. 222–229.

[10] K. Yasaswini, K. Puranik, A. Hande, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, Iiitt@ dravidianlangtech-eacl2021: Transfer learning for offensive language detection in dravidian languages, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 187–194.

[11] M. Shaheen, S. M. Awan, N. Hussain, Z. A. Gondal, Sentiment analysis on mobile phone reviews using supervised learning techniques, International Journal of Modern Education and Computer Science 11 (2019) 32.

[12] M. Anusha, F. Balouchzahi, H. Shashirekha, G. Sidorov, Mucic@ lt-edi-acl2022: Hope speech detection using data re-sampling and 1d conv-lstm, LTEDI 2022 (2022) 161.

[13] B. R. Chakravarthi, A. K. M, J. P. McCrae, B. Premjith, K. Soman, T. Mandl, Overview of the track on hasoc-offensive language identification-dravidiancodemix., in: FIRE (Working notes), 2020, pp. 112–120.

[14] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, D. García-Baena, J. García-Díaz, 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, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 378–388. URL: https://aclanthology.org/2022.ltedi-1.58. doi:10.18653/v1/2022.ltedi-1.58.

[15] 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.

[16] S. H. Lakshmaiah, F. Balouchzahi, A. M. Devadas, G. Sidorov, Coli-machine learning approaches for code-mixed language identification at the word level in kannada-english texts. acta polytechnica hungarica (2022).

[17] J. Armenta-Segura, G. Sidorov, A baseline for anime success prediction, based on synopsis, in: Congreso Mexicano de Inteligencia Artificial de la Sociedad Mexicana de Inteligencia Artificial COMIA-MICAI 2023 (Under Review), Zapopan, Jalisco, 2023.

[18] P. Soucy, G. W. Mineau, A simple knn algorithm for text categorization, in: Proceedings 2001 IEEE international conference on data mining, IEEE, 2001, pp. 647–648.

[19] F. Balouchzahi, H. Shashirekha, Mucs@ dravidian-codemix-fire2020: Saco-sentimentsanalysis for codemix text., in: FIRE (Working Notes), 2020, pp. 495–502.

[20] 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.

[21] S. Dowlagar, R. Mamidi, Edione@ lt-edi-eacl2021: Pre-trained transformers with convolutional neural networks for hope speech detection., in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 86–91.

[22] 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.

[23] I. S. Upadhyay, A. Wadhawan, R. Mamidi, et al., Hopeful_men@ lt-edi-eacl2021: Hope speech detection using indic transliteration and transformers, arXiv preprint arXiv:2102.12082 (2021).

[24] M. Shahiki Tash, Z. Ahani, A. Tonja, M. Gemeda, N. Hussain, O. Kolesnikova, Word level language identification in code-mixed Kannada-English texts using traditional machine learning algorithms, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 25–28. URL: https://aclanthology.org/2022.icon-wlli.5.

[25] 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.

[26] 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).

[27] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27, Springer, 2005, pp. 345–359.