ABCD Team at HOPE 2024: Hope Detection with BERTology Models and Data Augmentation

Bui Hong Son[1,2,*], Le Minh Quan[1,2,*] and Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents our participation in the HOPE tasks at IberLEF 2024[1, 2, 3, 4, 5], focusing on two of them: Task 1: Hope for Equality, Diversity, and Inclusion, and Task 2: Hope as Expectations. To address Task 1, we implemented and investigated different techniques and strategies. We first investigated the effectiveness of pre-processing steps for social media texts. Second, we employed two data augmentation strategies to tackle the class imbalance issue in the training dataset. Finally, we implemented a fine-tuning approach based on pre-trained language models combined with a simple ensemble technique. The private test results show that our best system achieved a top 5 ranking in Task 1. For Task 2, we achieved 2nd place in the binary classification subtask for Spanish datasets and 1st place for the same subtask on English datasets. Furthermore, our best results ranked 1st in the multi-classification subtask for both languages in the competition.

Keywords

Hope classification, Spanish language, English language, sentiment analysis, aspect-based sentiment analysis

1. Introduction

HOPE at IberLEF 2024 [1, 2, 3, 4, 5] is a competition that aims to analyze the multifaceted concept of hope through Natural Language Processing (NLP). HOPE shared-task consists of two different tasks for Equality, Diversity, and Inclusion. This task is to identify the messages that promote hope and acceptance for marginalized groups on social media platforms. The challenge is designed for competitors to develop various NLP models capable of differentiating between messages that uplift and empower these communities. Success hinges on your model's ability to accurately detect hope-oriented messages within this specific social media context. Task 2 - Hope as Expectations. This second task focuses on hope as it relate to future expectations and desires. The challenge here is to build NLP models proficient in detecting expressions of hope within social media text. These models need to not only identify hope, but also categorize its nature, distinguishing between realistic and unrealistic aspirations, as well as positive hope for

IberLEF 2024, September 2024, Valladolid, Spain

[*]Corresponding author.

22521246@gm.uit.edu.vn (B. H. Son); 20520709@gm.uit.edu.vn (L. M. Quan); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0000-0001-8340-1405 (D. V. Thin)

the future. Participating in HOPE 2024 is a unique opportunity to advance NLP significantly and address complex problems with real-world impact, pushing the boundaries of NLP tools and enhancing understanding of hope, social media, and human behavior.

In the previous year, HOPE at IberLEF 2023 [6] is also organized and focusing on the task of "Multilingual Hope Speech Detection" Various approaches were proposed and made public by numerous author. Among these, I2C-Huelva [7] Team applied a transformer model proposed for Spanish language, BERTuit. This team then achieved the second position and the first position for Spanish subtask and English subtask respectively. The same main approach is used by NLP URJC [8]. There is a little difference while this team applied BERT for English subtask and BETO for Spanish subtasks. With their optained results, they would have ranked 8th for the Spanish subtask and 1st for the English one. However, they missed the deadline for the paper submission. Distinct from the two preceding teams, besides testing XLM-R with different model setups, Zootopi Team [9] proposed two prompting scenarios for Large Language Model (ChatGPT) for the English and Spanish subtasks respectively. In the end, they achieved the 1st position in the Spanish subtask and ranked 9th in the English subtask. As we supposed, transformer-based models have been used in both subtasks and majority of the results are at the top of the competion's leaderboard. We cannot conclude that using tranformer-based models resulted in the better result than other approaches, such as traditional machine learning techniques like KNN (used by Zavira team [10]) or CNN (used be LIDOMA Team [11]), nor using the ChatGPT as Zootopi Team applied.

About the dataset for each task. In terms of Task 1, the dataset was collected between 2020 and 2023. It is an improved and extended version of the SpanishHopeEDI dataset [2]. The version of the dataset for IberLEF 2024 consists of training and dev sets on LGTB-related tweets and a test set on tweets related to the LGTBI collective and other EDI topics (unknown domains). A tweet is considered as HS if the text:

- i) explicitly supports the social integration of minorities;
- ii) is a positive inspiration for the collective;
- iii) explicitly encourages people who might find themselves in a situation;
- iv) unconditionally promotes tolerance

On the contrary, a tweet is marked as NHS if the text:

- i) expresses negative sentiment towards a community
- ii) explicitly seeks violence
- iii) uses gender-based insults

The dataset is composed of 2,000 tweets.

In terms of Task 2, the data collection commenced by retrieving the most recent 50,000 tweets between January and June 2022. Following this, an additional batch of 50,000 tweets was acquired within the same temporal scope using keywords associated with sentiments of hope. The dataset encompassed English and Spanish tweets originating from the first half of 2022, amounting to an aggregate of approximately 100,000 tweets per language.

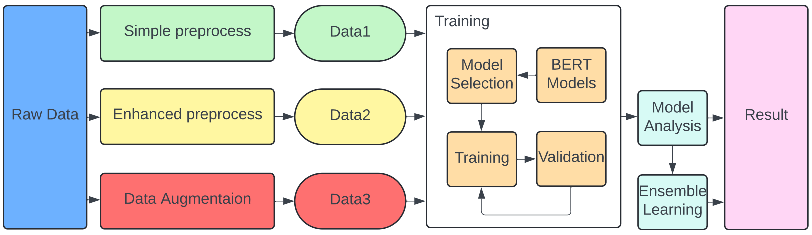
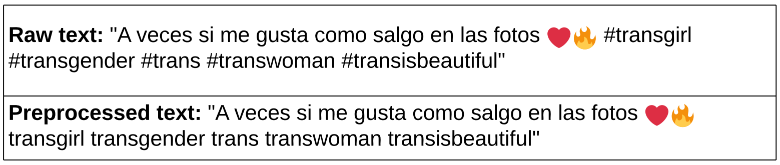


Figure 1: Our overall pipeline for the HOPE 2024 shared task.



Raw text: "A veces si me gusta como salgo en las fotos ❤️🔥 #transgirl #transgender #trans #transwoman #transisbeautiful"

Preprocessed text: "A veces si me gusta como salgo en las fotos ❤️🔥 transgirl transgender trans transwoman transisbeautiful"

Figure 2: Simple pre-processing sample.

2. Methodology

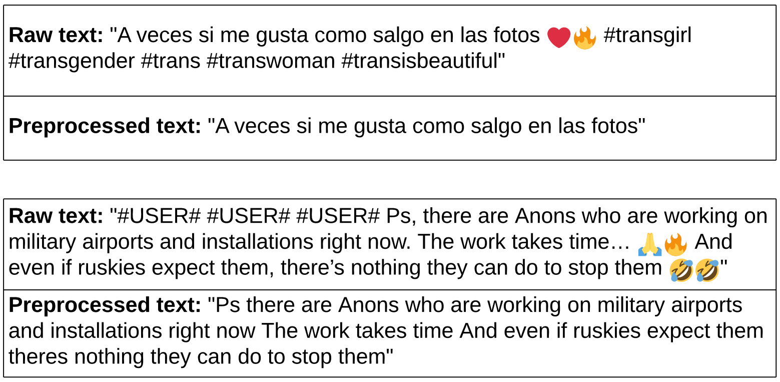
To address this challenge, we employ fine-tuning with different pre-trained language models for two tasks. We also investigate how pre-processing steps affect the models' performance. This is because the data originates from a social media platform, where proper pre-processing can significantly improve overall performance. Furthermore, we utilize various data augmentation techniques to enrich the training data. Finally, we implement a simple ensemble strategy to enhance performance for both tasks further. Figure 1 illustrates our overall pipeline for the HOPE 2024 shared task. Details of our main components are presented below.

2.1. Pre-processing Component

While analyzing the data, we discovered that the dataset contained noise and inconsistencies. To address this, applying pre-processing steps helped clean and standardize the data. This allowed the models to understand and context better, ultimately leading to more accurate results. To demonstrate the importance of pre-processing steps, we compare two strategies, including simple and specific strategies. We apply this method to both Task 1 and Task 2 to determine whether these pre-processing methods improve performance.

- **Simple pre-processing steps**: For this strategy, we only apply whitespace handling and punctuation removal. Figure 2 illustrates the steps in the simple pre-processing strategy.
- **Specific pre-processing steps**: For this strategy, we leverage the tweet-processer library[1]

[1] https://pypi.org/project/tweet-preprocessor/



Raw text: "A veces si me gusta como salgo en las fotos ❤️🔥 #transgirl #transgender #trans #transwoman #transisbeautiful"

Preprocessed text: "A veces si me gusta como salgo en las fotos"

Raw text: "#USER# #USER# #USER# Ps, there are Anons who are working on military airports and installations right now. The work takes time… 🙏🔥 And even if ruskies expect them, there's nothing they can do to stop them 🤣🤣"

Preprocessed text: "Ps there are Anons who are working on military airports and installations right now The work takes time And even if ruskies expect them theres nothing they can do to stop them"

Figure 3: Specific pre-processing steps.

because this library offers pre-processing functionalities that include: Emoji Removal, Username Removal, Specific Substring Removal, Hyperlink Removal, Text Normalization. Figure 3 shows an example of the specific pre-processing strategy.

2.2. Data Augmentation

We observed an imbalance issue between classes in Task 2. To improve overall performance, we aimed to expand the training data. To achieve this, we applied two data augmentation strategies to create new samples that can help the model learn more robust features. We briefly introduce two strategies which only applied for Task 2 as below:

- **Data Combination**: In this method, we combine the training datasets for English and Spanish into a single final dataset. We employ this strategy because we are utilizing multilingual models as our primary classifiers. Combining the datasets increases the number of data samples and leverages the strengths of multilingual language models.
- **Data Augmentation through Large Language Model**: Our main idea for this approach is to utilize the power of a pre-trained large language model to diversify the samples for imbalance classes. This work uses the Gemini models to create new samples through the prompt engineering with API function [2]. We send a request via API to run iterates through each text sample of the train set. With each sample, we order the Gemini to generate a distinct text with the same language and structure while still maintaining the expressiveness of the text.

[2] https://ai.google.dev/gemini-api/docs/api-overview?hl=vi

2.3. Classification Model

The Hope shared task[3] consists of two sub-tasks: Task 1: Hope for Equality, Diversity and Inclusion, and Task 2: Hope as Expectations. These sub-tasks involve binary classification and multi-class classification problems, respectively. To address these different tasks, we employ fine-tuning based on the pre-trained BERTology language models. Since several pre-trained language models support both English and Spanish, we implemented various models to investigate their performance on this shared task. A brief description of the models is presented below.

- **XLM-R** (Conneau et al. [12]): This powerful language model tackles tasks across 100 languages. It leverages a technique called self-supervised learning, where it analyzes a massive dataset (2.5TB of filtered CommonCrawl data) without any human intervention. This allows XLM-RoBERTa to learn from vast amounts of publicly available text, using an automated process to create both training examples and labels from the raw data itself. In this competition, we used both XLM-R-base and XLM-R-large.
- **DeBERTa** (He et al. [13]): We applied DeBERTa-v3-base, an improved version of DeBERTa in order to verify whether we get a superior result while using DeBERTa, a transformer-based neural language model designed to improve the BERT and RoBERTa models with two techniques: a disentangled attention mechanism and an enhanced mask decoder.
- **mDeBERTa-v3** (He et al. [14]): Building upon the success of DeBERTa, mDeBERTa V3 extends its capabilities to handle multiple languages. It retains the core structure of DeBERTa but leverages a massive dataset known as CC100, containing 2.5 trillion words across 100 languages. This base version boasts 12 processing layers and a hidden size of 768, allowing it to capture complex relationships within text. While the model itself has 86 million parameters, the vocabulary (the set of words it understands) adds another 190 million. This extensive vocabulary ensures that mDeBERTa V3 can effectively handle a vast range of languages.
- **RoBERTuito** (Pérez et al. [15]): A pre-trained model used for Sentiment Analysis in Spanish, used 500 milion tweets while training with the RoBERTa guidelines. RoBERTuito comes in 3 flavors: cased, uncased, and uncased+deaccented. In our experiments, we use base model.
- **Twitter-roBERTa** (Barbieri et al. [16]): This RoBERTa-base model specializes in understanding the sentiment of English tweets. Trained on a massive dataset of 58 million tweets, it can effectively analyze the emotions conveyed in social media messages. (Tweet-Eval benchmark used).
- **Twitter-XLM-roBERTa** (Barbieri et al. [17]): This XLM-RoBERTa model goes beyond just English. Trained on nearly 200 million tweets in eight languages (Arabic, English, French, German, Hindi, Italian, Spanish, and Portuguese), it can identify positive, negative, or neutral sentiment in social media posts. While it's pre-trained in these specific languages, it may even understand the sentiment in others. We decided to use this model to check whether it is effective while classifying different labels of social media texts.

[3] https://codalab.lisn.upsaclay.fr/competitions/17714

Table 1
The distribution of experimental datasets.

Labels	Training set	Validation set	Test set
Hope Speech (hs)	700	100	-
Not Hope Speech (nhs)	700	100	-
Total	1400	200	400

- **Bertin-RoBERTa** ([18]): A series of BERT-base models for Spanish text. We applied this model in order to observe if this model is better than traditional BERT on specific Spanish subtasks.

2.4. Ensemble Learning approach

To improve the overall performance of our models for the HOPE at IberLEF 2024 shared task, we leverage a max voting ensemble method. This technique is commonly used for classification tasks, which aligns well with the binary and multi-class classification problems in Hope's subtasks. In max voting, multiple models make predictions for each data point in the test set. Each model's prediction is considered a "vote," and the final prediction is the class label that receives the most votes from the ensemble.

3. Experimental Setup

3.1. Datasets and Evaluation Metrics

3.1.1. Task 1: Hope for Equality, Diversity and Inclusion

For Task 1, we used the official datasets provided by the organizers to train our models. To facilitate a comprehensive understanding of the data, we present both a table outlining the data distribution and a diagram illustrating the sequence lengths.

Table 1 presents the data distribution for the datasets used in Task 1. Divided into a training set (1400 samples), a validation set (200 samples) and a test set (400 samples). The data concerns classifying "Hope Speech" (hs) and "Not Hope Speech" (nhs). A balanced distribution is evident in the training set (700 samples each for hs and nhs), the same as the validation set (100 samples for each category).The data in the table indicates that all hope classes have a comparable number of participants (balanced). However, distribution across the three groups is uneven (different distribution variations). These balances play a crucial role in training and fine-tuning our models while also facilitating the resolution of data-related issues.

Besides, Figure 2 depicts the distribution of sequence length, that is, the number of words within a sequence, for two distinct categories in the datasets. There appear to be two distinct clusters of data points, suggesting a possible separation between the sequence lengths of "Hope Speech" and "Not Hope Speech" samples. Overall, the sequence length distributions for both categories exhibit a remarkable degree of similarity. However, the "hs" category appears to have some samples which have shorter sequences. The other category, "nhs" exhibits a broader distribution, encompassing a wider range of sequence lengths, including a small amount of longer samples.

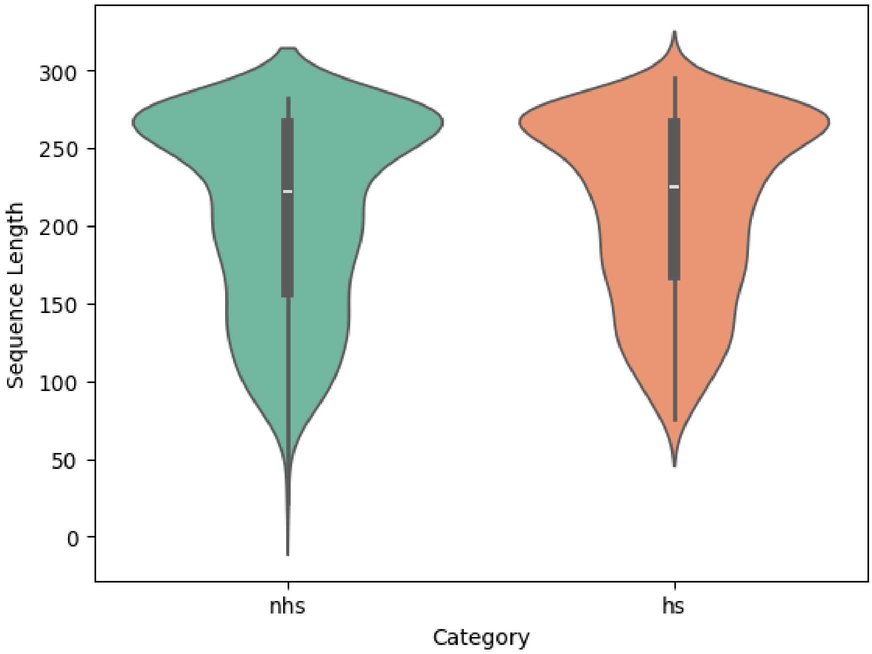


Figure 4: The distribution of sample length for each class in the training and validation sets.

3.1.2. Task 2: Hope as Expectations

In Task 2, we also use the original datasets provided by the organizers to train our models. Table 2 describes the statistics of datasets for Task 2.

Table 2
Statistics of official datasets for Task 2.

Type of labels		Spanish			English		
Binary	**multi-class**	**Train set**	**Validation set**	**Test set**	**Train set**	**Validation set**	**Test set**
Not Hope	Not Hope	4701	799	-	3088	502	-
Hope	Generalized Hope	1151	186	-	1726	300	-
	Unrealistic Hope	546	91	-	648	102	-
	Realistic Hope	505	74	-	730	128	-
	Total	6903	1150	1152	6192	1032	1032

As shown in Table 2, it can be seen that the distribution of data cross training, validation and test sets for binary and multi-class classification subtask in this Task. The hope can be categorized as either Binary (Hope or Not Hope) or multi-class (Not Hope, Generalized Hope, Unrealistic Hope, or Realistic Hope). The table separates the data into three sets: Train, Validation, and Test, showcasing how many instances of each sentiment label are included in each set.

In terms of the Spanish corpus, the data is imbalanced across the categories. For both binary and multi-class classifications, there are significantly more instances of Not Hope compared to the positive sentiment labels ("Hope" in Binary and "Generalized Hope", "Unrealistic Hope", and

"Realistic Hope" in multi-class). The imbalanced nature of the data can make it difficult for our model to learn the positive sentiment label accurately. The model might become biased towards the majority class ("Not Hope") and misclassify positive sentiment instances.

3.2. System Settings

We deployed our main framework with the support of the Hugging Face Transformer library. All models was set up to train with 10 epochs and the learning rate was set to 2e-4 for base models and 5e-5 for large models. Considering the size of the pre-trained language models, we chose a batch size of either 32 or 16. The hyperparameters of models are tuned based on the validation set. The majority of our training are trained on Kaggle, and the P100 accelerator was selected to accelerate our training. In terms of the tokenizer, in both tasks, we used the AutoTokenizer from the pre-trained model we imported from HuggingFace. The maximum length for the sequence that the Tokenizer will generate is 512. For all our experiments, we set a fixed random seed of 42 to train the models in both Task 1 and Task 2 (English datasets and Spanish datasets). The datasets used in Task 2 have two different languages, Spanish and English. However, we decided to apply the same pre-processing methods to all datasets. However, the pre-processing process included one of our main approaches in the experiments which is discussed it more later.

4. Experiment Results and Discussion

4.1. Task 1: Hope for Equality, Diversity and Inclusion.

In Task 1, we will observe and evaluate whether diversifying the provided datasets improves the final results. Also, we investigate whether Ensemble Learning results in different or improved results compared to the base model. Table 3 depicts the performance of four machine learning models (XLM-R-base, RoBERTuito, DeBERTa-v3-base, mDeBERTa-v3-base) on simple preprocessed-datasets and repeat 2 models (XLM-R-base, mDeBERTa-v3-base) on specific preprocessed-datasets.

Table 3
Experimental result Task 1: Hope for Equality, Diversity and Inclusion.

Datasets	Model	Avg. Macro F1	hs(a)			nhs(b)		
			Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
Simple pre-processing	XLM-R-base	58.79%	80.95%	80.95%	62.58%	73.33%	73.33%	55.00%
	RoBERTuito	54.81%	73.68%	73.68%	63.64%	49.43%	49.43%	45.99%
	DeBERTa-v3-base	56.06%	78.46%	78.46%	61.82%	62.69%	62.69%	50.30%
	mDeBERTa-v3-base	59.30%	85.00%	85.00%	63.57%	64.00%	64.00%	54.86%
Specific pre-processing	XLM-R-base	60.54%	74.36%	74.36%	65.17%	60.47%	60.47%	55.91%
	mDeBERTa-v3-base	60.26%	75.31%	75.31%	67.40%	61.04%	61.04%	53.11%
Ensemble Learning - Max Voting		**61.11%**	**82.35%**	**82.35%**	**66.67%**	**62.50%**	**62.50%**	**55.56%**

When trained on data with a simple pre-processing function, a metric used to evaluate models, at 56.06%. Other models performed with scores ranging from 48.79% to 54.81%. Remarkably, both models, XLM-R-base and mDeBERTa-v3-base, exhibited a significant improvement when trained on the dataset with specific pre-processing. The mDeBERTa-v3-base model archived a massive Macro F1-score in this scenario, reaching 60.54% in terms of F1-score. The remaining models

Table 4
Experimental result of Subtask 2.a: Binary Hope Speech Detection from Spanish datasets.

Datasets	Model	Spanish						
		M_Pr	M_Re	**M_F1**	W_Pr	W_Re	W_F1	Acc
Simple preprocess	XLM-R-base	82.66%	84.24%	**83.32%**	85.47%	84.90%	85.07%	84.90%
	RoBERTuito	83.03%	83.83%	**83.40%**	85.38%	85.16%	85.25%	85.16%
	Bertin-RoBERTa	82.91%	79.50%	**80.79%**	83.62%	83.85%	83.41%	83.85%
	Twitter-XLM-roBERTa	63.06%	84.03%	**83.61%**	85.63%	85.24%	85.38%	85.24%
	mDeBERTa	82.96%	84.30%	**83.54%**	85.60%	85.16%	85.30%	85.16%
Specific preprocess	RoBERTuito	82.88%	84.03%	**83.39%**	85.43%	85.07%	85.20%	85.07%
Combine 2 languages of datasets	RoBERTuito	81.16%	82.80%	**81.83%**	84.18%	83.51%	83.72%	83.51%
Generate data using AI	RoBERTuito	81.54%	83.06%	**82.17%**	84.44%	83.85%	84.04%	83.85%
Ensemble Learning - Max Voting		**83.69%**	**84.55%**	**84.09%**	**86.00%**	**85.76%**	**85.85%**	**85.76%**

also witnessed improvements, with scores ranging from 59.30% to 60.26%. Also, Ensemble learning significantly improves the overall accuracy of the classification, achieving an average Macro F1 score of 61.11%, the highest among all evaluated models.

These findings suggest that applying a wider range of pre-processing techniques can significantly enhance the performance of sentiment analysis models on social media data. While the DeBERTa-v3-base model achieved the highest with simple pre-processing, All models exhibited performance gains thanks to the enhanced dataset with additional processing steps.

Besides, we explore the application of Ensemble learning, especially Max Voting, to enhance the performance of sentiment analysis models for social media data. Our findings demonstrate that while the individual metrics for some models remain suboptimal, they still exhibit improvement compared to several single models. These results underscore the effectiveness of ensemble learning in boosting sentiment analysis performance and highlight the potential for further optimization through more sophisticated techniques.

4.2. Task 2: Hope as expectations

To inform the experimental result for Task 2, we have 4 tables. Table 4 and Table 5 represent the experimental results of binary classification task on both Spanish and English datasets, while Table 6 and Table 7 describe the result on English datasets.

4.2.1. Subtask 2.a: Binary Hope Speech Detection

Table 4 presents the experimental findings for Task 2 Binary on Spanish datasets. Among the models trained on the simply preprocessed dataset, Twitter-XLM-roBERTa achieved the best performance with an M_F1 score of 83.61%. However, we decided to utilize the RoBERTuito model for further experiments in this task as it is specifically trained for Spanish social media data. However, despite employing more approaches, the subsequent methods failed to result in any improvements. Finally, only by implementing Ensemble Learning based on the previously obtained results did we observe a significant improvement and achieve the highest M_F1 score of 84.09%.

Table 5 depicts the influence of various techniques on the performance of our BERT models. Among the evaluated models, the XLM-R-base exhibited the most promising performance on the basic dataset with simple pre-processing, achieving the highest F1-score of 86.63%. The

Table 5
Experimental result of Subtask 2.a: Binary Hope Speech Detection from English datasets.

Datasets	Model	English						
		M_Pr	M_Re	**M_F1**	W_Pr	W_Re	W_F1	Acc
Simple preprocess	**XLM-R-base**	**86.65%**	**86.53%**	**86.58%**	**86.64%**	**86.63%**	**86.62%**	**86.63%**
	DeBERTa-v3-base	85.62%	85.15%	**85.26%**	85.52%	85.37%	85.32%	85.37%
	Twitter-roBERTa	84.65%	84.31%	**84.40%**	84.58%	84.50%	84.46%	84.50%
	Twitter-XLM-roBERTa	85.34%	84.59%	**84.73%**	85.20%	84.88%	84.80%	84.88%
Specific preprocess	XLM-R-base	86.08%	85.94%	**85.99%**	86.06%	86.05%	86.03%	86.05%
Combine 2 languages of datasets	XLM-R-base	85.76%	85.36%	**85.46%**	85.68%	85.56%	85.52%	85.56%
Generate data using AI	XLM-R-base	83.25%	83.32%	**83.23%**	83.37%	83.24%	83.25%	83.25%
Ensemble Learning: Max Voting		86.46%	86.18%	**86.26%**	86.40%	86.34%	86.31%	86.34%

Table 6
Experimental result of Subtask 2.b: multi-class Hope Speech Detection from Spanish datasets.

Datasets	Model	Spanish						
		M_Pr	M_Re	**M_F1**	W_Pr	W_Re	W_F1	Acc
Simple preprocess	XLM-R-base	64.42%	66.55%	**65.29%**	81.57%	80.64%	81.03%	80.64%
	RoBERTuito	62.41%	58.95%	**60.49%**	77.58%	78.65%	78.02%	78.65%
	Bertin-RoBERTa	65.64%	62.83%	**64.09%**	79.92%	80.56%	80.16%	80.56%
	Twitter-XLM-roBERTa	61.69%	65.43%	**63.21%**	80.85%	78.47%	79.34%	78.47%
	mDeBERTa	62.63%	66.86%	**64.41%**	80.79%	78.91%	79.70%	78.91%
Specific preprocess	Bertin-RoBERTa	61.59%	60.38%	**60.94%**	78.56%	78.82%	78.65%	78.82%
Combine 2 languages of datasets	Bertin-RoBERTa	59.98%	59.33%	**59.38%**	77.69%	77.17%	77.26%	77.17%
Generate data using AI	Bertin-RoBERTa	64.93%	59.76%	**61.80%**	79.12%	80.21%	79.42%	80.21%
Ensemble Learning - Max Voting		**68.31%**	**65.43%**	**66.68%**	**81.80%**	**82.03%**	**81.78%**	**82.03%**

remaining models trained on the same datasets resulted in M_F1 scores ranging from 84.88% to 85.37%. Remarkably, applying additional pre-processing or data augmentation techniques did not resulted in any significant improvements for these models. In some cases, it even caused performance decreases compared to simple pre-processing scenarios. Besides, while Ensemble Learning did not achieve the absolute best results, it demonstrated a notable improvement compared to individual models' results.

4.2.2. Subtask 2.b: Multi-class Hope Speech Detection

As described in Table 6, the models performed well on the dataset subjected to basic pre-processing. Among these, XLM-R-base and Bertin-RoBERTa models achieved the highest and second-highest M_F1 scores of 65.29% and 64.09%, respectively. However, we decided to employ additional approaches on Bertin-RoBERTa to obtain more objective results using a model fine-tuned specifically for the Spanish texts. Consequently, methods such as Specific pre-processing, training the model using a combined train dataset, or generating more data did not cause any remarkable results, while applying the Max Voting ensemble technique resulted in the best performance, with an M_F1 score of 66.68%.

Table 7 presents the experimental results of Task 2 multi-class Classification on the English Dataset. Overall, DeBERTa-v3 resulted in a remarkable performance on the simple processed dataset with an M_F1 score of 69.92% compared to Twitter-XLM-RoBERTa with an M_F1 score of 69.00%. Nonetheless, we decided to choose Twitter-XLM-RoBERTa for further investigations because it is a pretrainned model for sentiment analysis with social media text. Upon com-

Table 7
Experimental result of Subtask 2.b: multi-class Hope Speech Detection from English datasets.

Datasets	Model	English						
		M_Pr	M_Re	**M_F1**	W_Pr	W_Re	W_F1	Acc
Simple preprocess	DeBERTa-v3-base	69.19%	70.80%	**69.92%**	76.33%	75.58%	75.89%	75.58%
	Twitter-XLM-roBERTa	68.27%	69.85%	**69.00%**	75.63%	75.10%	75.32%	75.10%
Specific preprocess	Twitter-XLM-roBERTa	68.60%	70.21%	**69.34%**	76.03%	75.39%	75.65%	75.39%
Combine 2 languages of datasets	Twitter-XLM-roBERTa	70.10%	72.42%	**71.09%**	77.52%	76.55%	76.92%	76.55%
	XLM-R-large	**71.38%**	**72.82%**	**72.00%**	**78.05%**	**77.42%**	**77.67%**	**77.42%**
Generate data using ChatBot	Twitter-XLM-roBERTa	66.31%	66.11%	**66.16%**	72.14%	72.38%	72.21%	72.38%
Ensemble Learning - Max Voting		71.03%	72.21%	**71.57%**	77.66%	77.13%	77.35%	77.13%

Table 8
Ranking of our systems on two sub-tasks.

Task 2	Datasets	Model/Method	Score							Ranking
			M_Pr	M_Re	M_F1	W_Pr	W_Re	W_F1	Acc	
Subtask 2.a: Binary Classification	Spanish	Ensemble Learning	83.69%	84.55%	**84.09%**	86.00%	85.76%	85.85%	85.76%	2
	English	XLM-R-base	86.65%	86.53%	**86.58%**	86.64%	86.63%	86.62%	86.63%	1
Subtask 2.b: Multiclass Classification	Spanish	Ensemble Learning	68.31%	65.43%	**66.68%**	81.80%	82.03%	81.78%	82.03%	1
	English	XLM-R-large	71.38%	72.82%	**72.00%**	78.05%	77.42%	77.67%	77.42%	1

bining two datasets of different languages, Twitter-XLM-RoBERTa exhibited improvement in performance, achieving an M_F1 score of 71.09%, an increase of 2.09% compared to the original one. The technique of generating additional data did not result in any noticeable enhancement, whereas Ensemble Learning caused significant improvement in the overall result, with an M_F1 score of 71.57%. Finally, with the assistance of ample resources, we were able to employ the XLM-RoBERTa-large model for experiments in this task. Naturally, utilizing the large model resulted in the highest performance, with an M_F1 score of 72.00%.

5. System Ranking

Concerning Task 1, among the employed models, Ensemble Learning ultimately resulted in the best Average Macro F1 score is 61.11%. However, the XLM-R-base model caused the highest Precision score, so we submitted its prediction achieved fifth place in the overall ranking with Average Macro F1 score is 58.79%.

In terms of Task 2, the official ranking results are presented in Table 8. For Task 2 - Subtask 2.a: Binary Hope Speech Detection from Spanish datasets, Ensemble Learning emerged as the most efficacious method, achieving an M_F1 score of 84.09%, which serves as the benchmark metric for ranking. Our system in this task attained the second position. For Task 2 - Subtask 2.a: Binary Hope Speech Detection from English datasets, we attained the first position with an M_F1 score of 86.58%, demonstrating a superior outcome compared to the preceding two tasks, leveraging the XLM-R model. Transitioning to Task 2 - Subtask 2.b: multi-class Hope Speech Detection from Spanish datasets, our methodology reached an M_F1 score of 66.68% and secured the best rank utilizing the Ensemble Learning technique. Finally, in the final Task - Subtask 2.b: multi-class Hope Speech Detection from English datasets, with a M_F1 score of 72.00%, we attained the topmost position employing the XLM-R model.

6. Conclusion

This work presented our system architecture, experimental procedures, and final ranking in the HOPE 2024 competition. We implemented various techniques to investigate the performance of this shared task. This included the simple and specific pre-processing steps, dataset combination across languages, and data augmentation with large language models. We rigorously evaluated these methodologies using pre-trained models for the sub-tasks. Finally, our approach achieved the top scores in various sub-tasks. Specifically, our best system ranked in the Top 5 for Task 1, Top 2 and Top 1 for Task 2 - PolyHope Binary (Spanish and English). For Task 2 - PolyHope multi-class, we reach the Top 1 for English and Spanish language.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] 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.

[2] 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).

[3] 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.

[4] 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.

[5] G. Sidorov, F. Balouchzahi, S. Butt, A. Gelbukh, Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets, Applied Sciences 13 (2023) 3983.

[6] S. M. Jiménez-Zafra, F. Rangel, M. M.-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), CEURWS. org, 2023.

[7] J. L. D. Olmedo, J. M. Vázquez, V. P. Álvarez, I2c-huelva at hope2023@ iberlef: Simple use of transformers for automatic hope speech detection (2023).

[8] M. Á. Rodríguez-García, A. Riaño-Martínez, S. M. Herranz, Urjc-team at hope2023@ iberlef: Multilingual hope speech detection using transformers architecture (2023).

[9] A. Ngo, H. T. H. Tran, Zootopi at hope2023iberlef: Is zero-shot chat-gpt the future of hope speech detection, 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), CEURWS. org, 2023.

[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] M. S. Tash, J. Armenta-Segura, O. Kolesnikova, G. Sidorov, A. F. Gelbukh, Lidoma at hope2023@iberlef: Hope speech detection using lexical features and convolutional neural networks, in: IberLEF@SEPLN, 2023. URL: https://api.semanticscholar.org/CorpusID:265309454.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[13] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2020.

[14] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[15] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, Robertuito: a pre-trained language model for social media text in spanish, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[16] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1644–1650. URL: https://aclanthology.org/2020.findings-emnlp.148. doi:10.18653/v1/2020.findings-emnlp.148.

[17] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[18] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.