Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 407-419 recibido 07-07-2024 revisado 17-07-2024 aceptado 30-07-2024

Overview of HOPE at IberLEF 2024: Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations

Resumen de la tarea HOPE en IberLEF 2024: Abordando la Detección del Discurso de Esperanza en los Medios Sociales desde Dos Perspectivas, para la Igualdad, la Diversidad y la Inclusión y como Expectativas

Daniel García-Baena,[1] **Fazlourrahman Balouchzahi,**[2] **Sabur Butt,**[3]
Miguel Ángel García-Cumbreras,[1] **Atnafu Lambebo Tonja,**[2] **José Antonio García-Díaz,**[4]
Selen Bozkurt,[5] **Bharathi Raja Chakravarthi,**[6] **Hector G. Ceballos,**[3]
Rafael Valencia-García,[4] **Grigori Sidorov,**[2] **L. Alfonso Ureña-López,**[1]
Alexander Gelbukh,[2] **Salud María Jiménez-Zafra**[1]
[1]Computer Science Department, SINAI, CEATIC, Universidad de Jaén (Spain)
daniel.gbaena@gmail.com, {magc,laurena,sjzafra}@ujaen.es
[2]Centro de Investigación en Computación, IPN (Mexico)
{fbalouchzahi2021,alambedot2022,sidorov,gelbukh}@cic.ipn.mx
[3]Institute for the Future of Education (IFE) at Tecnológico de Monterrey (Mexico)
{saburb,ceballos}@tec.mx
[4]Facultad de Informática, UMUTeam, Universidad de Murcia (Spain)
{joseantonio.garcia8,valencia}@um.es
[5]Department of Biomedical Informatics, School of Medicine, Emory University
selen.bozkurt@emory.edu
[6]University of Galway, Ireland, bharathi.raja@insight-centre.org

Abstract: This paper presents the second edition of the international shared task on multilingual hope speech detection, HOPE 2024, conducted as part of the IberLEF workshop during the SEPLN 2024 conference. This shared task encompassed two distinct subtasks: the detection of hope speech within Equality, Diversity, and Inclusion texts and the identification of hope speech focusing on expectations in two-level, binary, and multiclass classification settings. Nineteen teams participated in the competition, and sixteen submitted their working notes. In the first subtask, the top-ranking team achieved an average Macro F1-score of 71.61. In the second subtask, leading teams demonstrated robust performance with F1 scores exceeding 80.00 for binary classification and 78.00 for multiclass classification settings.
Keywords: Hope speech, LGBTQ, EDI, NLP.

Resumen: Este artículo presenta la segunda edición de la tarea compartida internacional sobre detección multilingüe del discurso de la esperanza, HOPE 2024, desarrollada como parte del taller IberLEF durante el congreso SEPLN 2024. Esta tarea estuvo compuesta por dos subtareas distintas: la primera de ellas se centró en la detección del discurso esperanzador en el ámbito de la Igualdad, Diversidad e Inclusión, mientras que la segunda se enfocó en las expectativas del discurso esperanzador atendiendo a dos niveles, binario y multiclase. Diecinueve equipos participaron en la competición y dieciséis enviaron sus notas de trabajo. En la primera subtarea, el equipo mejor clasificado logró una puntuación Macro F1 media de 71,61. En la segunda subtarea, los equipos líderes demostraron un rendimiento sólido con puntuaciones F1 superiores a 80,00 para la clasificación binaria y 78,00 para los entornos de clasificación multiclase.
Palabras clave: Discurso esperanzador, LGBTQ, DEIB, PLN.

ISSN 1135-5948 DOI 10.26342/2024-73-31

1 Introduction

Hope is one of the exceptional human capacities that allows for flexible anticipation of future events and possible expected outcomes. These visions significantly influence emotions, behaviour, and mood (Bruininks and Malle, 2005). Individuals with high hope do not react in the same way to barriers as individuals with low hope, but instead view barriers as challenges to overcome and use their pathway thoughts to plan an alternative route to their goals (Snyder, 1994; Snyder, 2002). In addition, high hope has been found to correlate with a number of beneficial elements, such as academic performance (Snyder, 2002) and lower levels of depression (Snyder et al., 1997). In contrast, low hope is associated with negative outcomes, such as reduced well-being (Eid and Larsen, 2008).

Despite the importance and prevalence of hope, it has received little attention in the field of Natural Language Processing (NLP). Machine learning and NLP techniques can be used to analyze social media data and provide insights into the nature of hope in human behavior and decision-making. Therefore, in the last two years, we have promoted research on this topic by organizing shared tasks. We organized shared tasks on hope speech detection at the second workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-2022), as a part of ACL 2022 (Chakravarthi et al., 2022), at LT-EDI-2023, within RANLP 2023 (Kumaresan et al., 2023) and the shared task HOPE in IberLEF 2023 (Jiménez-Zafra et al., 2023). The participants registered in these previous tasks show that there is an expected target community, as the maximum number of participants registered was 126 and the minimum 50.

The main novelty of this new edition is the study of hope from two perspectives: i) hope for equality, diversity and inclusion, and ii) hope as expectations. The first perspective was explored in the last edition of IberLEF 2023 for English and Spanish. In this new edition, we focus on expanding and improving our Spanish dataset to answer one of the most frequently asked questions from previous participants and researchers in hope speech detection: Is it possible to detect hope speech in multiple domains, even when we only train our models with texts from one specific area? Therefore, this time we again provide the teams with a training corpus focused on the LGTBI community, but we ask them to test their systems with texts belonging to the LGTBI domain and new unknown domains. The second perspective has not been studied previously in any shared task and, for the IberLEF 2024 edition (Chiruzzo, Jiménez-Zafra, and Rangel, 2024), we propose its study from a binary and multi-class perspective for English and Spanish.

The rest of the paper is organized as follows. In Section 2, the task is described and the provided datasets are presented in Section 3. Then, Section 4 explains how the task was set up, and Section 5 describes the participants' approaches. Subsequently, Section 6 presents the results obtained by the participants and a discussion of the same. Finally, the conclusions of the task and future directions are shown in Section 7.

2 Task description

The aim of the HOPE 2024 shared task is to promote Equality, Diversity and Inclusion (EDI) through the detection of hope speech and analyze the expectations of this type of speech. The general challenges proposed for this second edition are:

1. To explore different Machine Learning approaches for multiclass and multilingual hope detection.
2. To experiment with techniques dealing with the imbalanced distribution of labels.
3. To experiment with noisy social media data for hope detection.
4. To detect hope speech in domains other than those in which our systems have been trained.
5. To address the linguistic differences between English and Spanish texts in the realm of hope expressions.
6. To deal with a lack of context while identifying hope from a text.
7. To distinguish realistic and unrealistic as well as general hope/expectation computational.

This shared task is divided into two subtasks, which are described below.

2.1 Subtask 1: Hope for Equality, Diversity and Inclusion

Hope speech is the type of speech that is able to relax a hostile environment (Palakodety, KhudaBukhsh, and Carbonell, 2019) and that helps, gives suggestions and inspires for good to a number of people when they are in times of illness, stress, loneliness or depression (Chakravarthi, 2020). On social media, offensive messages are posted against people because of their race, color, ethnicity, gender, sexual orientation, nationality, or religion. As Chakravarthi (2020) stated, how vulnerable groups interact with social media has been studied and found that it plays an essential role in shaping the individual's personality and view of society (Burnap et al., 2017; Kitzie, 2018; Milne et al., 2016). Examples of these vulnerable groups are the Lesbian, Gay, Bisexual, and Transgender (LGBT) community, racial minorities or people with disabilities. This subtask is related to the inclusion of vulnerable groups and focuses on the study of the detection of hope speech, in pursuit of EDI. It consists of, given a tweet written in Spanish, identify whether it contains hope speech or not. Specifically, it is divided into two subtasks: subtask 1.a: Hope speech detection on LGTBI domain; and subtask 1.b: Hope speech detection on unknown domains. The possible categories for each text are:

- hs: Hope Speech.
- nhs: Non-Hope Speech.

2.2 Subtask 2: Hope as expectations

In this subtask, the challenge of hope speech detection was approached from a novel theoretical perspective. Hope was conceptualized as a future-oriented expectation regarding either a general or specific outcome. Consequently, the classification task was structured as a two-tiered hope speech detection process. It focuses on expectations, and desirable and undesirable facts. Specifically, it is divided into two subtasks: subtask 2.a: Binary Hope Speech Detection from English and Spanish texts; and subtask 2.b: Multiclass Hope Speech Detection from English and Spanish texts. Initially, texts expressing any form of expectation were identified as *Hope*, while the remaining texts were categorized as *Not Hope*. In the multiclass setting, the *Not Hope (NH)* category remained unchanged, whereas the *Hope* category was further refined into three subcategories: *General Hope (GH)*, *Realistic Hope (RH)*, and *Unrealistic Hope (URH)*. Detailed descriptions of each class are presented in Balouchzahi, Sidorov, and Gelbukh (2023) and summarized below.

- *Generalized Hope* - is expressed as a general hopefulness and optimism that is not directed toward any specific event or outcome.
- *Realistic Hope* - is about expecting something reasonable, meaningful, and possible thing to happen.
- *Unrealistic Hope* - is usually in the form of wishing for something to become true, even though the possibility of happening is remote or significantly less or even zero.
- *Not Hope* - texts that do not convey hope.

3 Datasets

For the first subtask, we retagged the texts from last year IberLEF shared task dataset, the SpanishHopeEDIv2 (Jiménez-Zafra et al., 2023), with some new domain-general annotation guidelines and collected a total of 817 texts for label hs and 833 for category nhs. In addition to retagging, we removed some duplicated posts from SpanishHopeEDIv2. The resulting distribution was of 791 hs and 821 nhs texts. Later on, and to maintain the balance between hs and nhs categories, we obtained some additional posts from the social network X and marked them in relation to their category: hs and nhs, and topic: we collected texts related to LGBT, obesity and racism. We added 9 from those new texts to subset of 791 hs texts and removed 21 from the 821 available posts from the nhs category, achieving a total of 800 entries in both cases. As a result, for the new SpanishHopeMultidomain corpus (see Table 1), we took 1,400 (700 hs and 700 nhs) texts for the training subset and 200 (100 hs and 100 nhs) for the development subset. The test subset was composed of a total of 400 texts (200 hs and 200 nhs). The topic of the texts in the training and development subsets was always LGBT, but for the test subset, 200 texts were LGBT-related (100 hs and

100 nhs), 106 were about obesity (54 hs and 52 nhs), and 94 were concerned with racism (46 hs and 48 nhs).

Type	Dev	Train	Test
Hope	100	700	200
Non-Hope	100	700	200
Total	**200**	**1,400**	**400**

Table 1: SpanishHopeMultidomain dataset statistics (Subtask 1).

On the other hand, for the second subtask, the data collection process commenced with the retrieval of the most recent 50,000 tweets from the X platform in English, spanning the period from January 2022 to June 2022. Subsequently, an additional batch of 50,000 tweets was obtained within the same timeframe, utilizing keywords associated with sentiments of hope. These keywords included terms such as "hope," "Inshallah," "aspire," "believe," "expect," "want," and "wish," along with their various forms and synonyms.

To extend the dataset to Spanish, the same keywords were translated into Spanish and rigorously validated by native speakers. This process ensured the comprehensive inclusion of any contextually relevant terms that might have been overlooked, thereby creating the Spanish version of the PolyHope dataset (Balouchzahi, Sidorov, and Gelbukh, 2023). The data collection for Spanish adhered to the same methodology as for English, ultimately yielding 8,033 labeled tweets, categorized according to different aspects of hope.

The detailed statistics for the datasets pertaining to the second subtask are presented in Table 2. This table illustrates the distribution of the tweets across various hope categories, providing an insightful overview of the dataset composition and the scope of sentiments captured within this study. Such meticulous data collection and validation processes are crucial for ensuring the robustness and reliability of the PolyHope dataset, facilitating subsequent analyses and research endeavors focused on the sentiment of hope across different languages and cultural contexts.

4 Task Settings

This shared-task was organized through CodaLab[1] and was divided into three stages: Practise, Evaluation, and Post-evaluation. These stages are described below.

4.1 Practice

During the practice phase, we provided participants with labelled training and development data that they could use to train and validate their models. We released the data for Spanish and English, depending on the subtask, so that participants could develop their systems for one or both subtasks. The objective of this first phase was to provide all teams with sufficient data for them to use in their preliminary evaluations and hyperparameter tuning. This ensured that participants were ready for evaluation prior to the release of the unlabeled test data. In this phase, both for subtask 1 and 2, participants were allowed to make a maximum of 100 submissions through CodaLab in order to know the performance of their systems.

4.2 Evaluation

On the evaluation phase participants received the test dataset without the gold labels. Each team could participate with up to 10 submissions for each subtask from which they had to select the best ones for the ranking. In subtasks 1.a and 1.b, the systems were evaluated using Macro-precision, Macro-recall and Macro-F1, and they were ranked using the average Macro-F1 score of both subtasks. On the other hand, in subtasks 2.a and 2.b, the systems were evaluated using the macro and weighted scores of precision, recall, and F1, and they were ranked using the Macro-F1 score.

4.3 Post-evaluation

After the evaluation phase, a post-evaluation phase was opened in which participants could test improved versions of their systems and in which new users can participate to test their approaches. This phase remains open and is still accessible.

5 Participant approaches

The HOPE 2024 shared task attracted 56 teams that registered through CodaLab, 23 results were submitted and 16 working notes

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

Language	Set	Binary		Multiclass				Total
		NH	HS	NH	GH	RH	URH	
English	Train set	3,088	3,104	3,088	1,726	730	648	6,192
	Dev set	502	530	502	300	128	102	1,032
	Test set	491	541	491	309	124	108	1,032
Spanish	Train set	4,701	2,202	4,701	1,151	505	546	6,903
	Dev set	799	351	799	186	74	91	1,150
	Test set	773	379	773	206	77	96	1,152

Table 2: Statistics of the dataset for Subtask 2 (HS: Hope Speech, NH: Not Hope, GH: General Hope, RH: Realistic Hope, URH: Unrealistic Hope).

describing the systems were presented. In this section, we briefly describe each of the proposals submitted by the participants.

5.1 NTT@UIT

In relation to subtask 1, the team NTT@UIT (Nguyen Thi and Dang Van, 2024), from the University of Information Technology-VNUHCM (Vietnam) and Vietnam National University, used ChatGPT 3.5 with three different prompting techniques: zero-shot, few-shot (one-shot, three-shot) and chain of thought, all of them combined with six different information strategies. As it was with Zootopi team from last year shared task (Ngo and Tran, 2023), the team that used ChatGPT for the automatically detecting hope speech was the best scorer this time too. Specifically, team NTT@UIT achieved the best average Macro-F1, 71.61, using prompting, which combined a zero-shot prompting approach while providing some information strategy. See Table 3 for the complete results.

The team used the same approach for the second sub-task and scored a micro-F1 73 in binary classification for both Spanish and English. They place in 9th position in multiclass tasks for both languages.

5.2 ChauPhamQuocHung

ChauPhamQuocHung (Chau Pham Quoc and Dang Van, 2024), from the University of Information Technology-VNUHCM (Vietnam) and Vietnam National University too, scored second for subtask 1 with an average Macro-F1 value of 65.79. The researchers tried three different LLMs, all available in Hugging Face, as they are: BETO (Cañete et al., 2023), RoBERTuito (Pérez et al., 2022) and SpanBERTa. Models were pre-trained on OPUS Project + Spanish Wikipedia, Spanish OSCAR, and Spanish tweets, respectively. They performed some pre-processing by converting emojis to text with the emoji package (`https://pypi.org/project/emoji/`) and removing URL links. They did not do any hyper-parameter optimization.

In subtask 2, the ChauPhamQuocHung team was placed in the 3rd place with a micro F1 scores of 83 and 85 in the binary classification tasks for Spanish and English, respectively. They used language-specific pre-trained models and multilingual models in their experiments.

5.3 CUFE

The team named CUFE (Ibrahim, 2024), from Cairo University (Egypt), was the only one that used Gzip compression for text classification. Their proposal, despite being way different from all the rest, performed unexpectedly well when it was compared with popular and widely tested LLMs as it is BERT. The member of team CUFE preprocessed all the texts, deleting non-alphanumeric symbols, URLs, HTML tags, punctuations, Spanish stop words, and emojis. They managed to achieve a very close average Macro-F1 to the one from the second-ranked team (ChauPhamQuocHung) with a value of 65.22.

5.4 Ometeotl

Ometeotl researchers (Armenta-Segura and Sidorov, 2024), the team from the Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC IPN) of Mexico, they proposed the use of a custom BERT-based model (as it is the one available in Hugging Face: nlptown/bert-base-multilingual-uncased-sentiment), that was specifically pre-trained on multilingual datasets and tailored for sentiment analysis. Even with an apparently simpler approach, they scored close to

third position in rank with an average Macro-F1 score of 64.38. See Table 3 for the complete results.

In subtask 2, Ometeotl team achieved 81 (5th) and 82 (5th) micro F1 score in binary classification task for Spanish and English respectively.

5.5 ABCD Team

ABCD Team (Bui Hong, Le Minh, and Dang Van, 2024), conformed by researchers from the University of Information Technology-VNUHCM (Vietnam) and Vietnam National University, as they were the first and the third teams in the rank from Table 3, they tested the performance of four machine learning models (XLM-R-base, RoBERTuito, DeBERTa-v3-base, mDeBERTa-v3-base) on simple pre-processed datasets (whitespace handling and punctuation removal) and repeated the same with another two models (XLM-R-base and mDeBERTa-v3-base) and some more specifically pre-processed datasets (using the tweet-processer library: `https://pypi.org/project/tweet-preprocessor/`). In addition, they explored the application of ensemble learning, especially the Max Voting technique, where they managed to achieve their best results, an average Macro-F1 value of 58.79.

For subtask 2 they wined in 2nd place with 84 micro-F1 score in binary classification task for Spanish and they wined the competition in 1st place in binary classification task for English with micro-F1 score 87. For multiclass classification task they obtained 1st position for both languages with 72 and 67 micro-F1 score for English and Spanish respectively.

5.6 MUCS

Team MUCS (Divakaran, Girish, and Shashirekha, 2024) (Mangalore University, India), they developed some preprocessing converting emojis to text using demoji library (`https://pypi.org/project/demoji/`) and converted all numeric information to words while removing URLs, user mentions, hash tags, special characters, punctuation and stopwords. They tested ML methods as: SVM, LR, LSVC, RF, CatBoost, XGBoost and AdaBoost. In addition, they tried using some TL (transfer learning) with DistilSpanBERT and mBERT. Finally, they tried with an ensemble model called: Hope_probfuse, based in soft voting. As a result, the researchers from MUCS scored an average Macro-F1 value close to 50.00 and, contrary to the rest of teams, did not obtain the best results while working with transformers.

MUCS team obtained 4th and 5th place in binary classification task for Spanish and English respectively with 82 micro-F1 score in both languages.

5.7 VEL

VEL team (Ponnusamy et al., 2024) was conformed by a compound of researchers from: Digital University Kerala, India, University of Galway, Ireland and Gandhigram Rural Institute-Deemed to be University, India. They tried with different ML models as some of the previous teams but it was the popular LLM BERT based pretrained model: bert-base-spanish-wwm-cased, from Hugging Face, the one that obtained the best average Macro-F1 results. This team, as most of the rest, made some preprocessing too. They achieved position 7th with an average Macro-F1 value of 51.03.

5.8 CICPAK

CICPAK team (Ullah et al., 2024), from Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico, used BERT for classifying all texts. This team made some preprocessing lowercasing, removing URL links and emojis, and removing punctuation symbols, numerical data, commonly occurring words, stopwords, and uninformative phrases. They scored an average Macro-F1 value of 49.48 (8th).

The CICPAK team used the same approach to solve both binary and multi-class tasks for subtask 2. They scored 71(12th) and 74 (8th) Micro-F1 in binary tasks for Spanish and English, respectively, and 45 (8th) and 30 (10th) Micro-F1 scores in multiclass tasks for Spanish and English.

5.9 UMUTeam

UMUTeam (Pan, Ángela Almela, and Alcaraz-Mármol, 2024), from Universidad de Murcia, Spain, they adopted an approach that involved fine-tuning different pre-trained models and adding later some additional features to the examples, such as emotions and sentiments extracted from text. This was

done by concatenating the obtained text embeddings with the outputs (probabilities) of emotion and sentiment identification models, as it was pysentimiento (Pérez et al., 2023). Specifically, they evaluated MarIA, BETO and BERTIN, achieving the best results while using MarIA. The researchers from UMUTeam scored 9th with an average Macro-F1 score of 48.90.

UMUTeam obtained 3rd and 2nd place in binary classification task for Spanish and English respectively with 83 and 86 micro-F1 score. For multi-class task they obtained 2nd and 5th place for English and Spanish task respectively with micro F1 of 68 and 63.

5.10 UniOfGalway

UniOfGalway (Subburaj et al., 2024) (University of Galway, Ireland), they made some data preprocessing, removing numbers, punctuation symbols, stopwords and emojis and lowercassing. This team made several experiments with a range of transformers models such as bert-base-spanish-wwm-cased-finetuned-spa-squad2-es, bert-base-spanish-wwm-cased, bert-base-spanish-wwm-uncased, xlm-roberta-base, roberta-base-bne and electricidad-base-discriminator. Getting the best Macro-F1 score values with: bert-base-spanish-wwm-cased. All models were taken from Hugging Face. They scored with an average Macro-F1 value of 45.66 (10th).

5.11 Grima

Girma team (Bade et al., 2024), they tested three different algorithms, including logistic regression, Word2Vec and some transformer-based models. This team made some data pre-processing, handling missing data and performing data cleaning and encoding. In both subtasks, they scored better results when they worked with transformers. They managed to achieve an 44.40 average F1-score (11th).

For task 2, the Girma team used the same algorithms as task 1 and performed different pre-processing techniques before training models. They scored 75 (10th) and 82 (5th) in the binary classification task in Spanish and English for sub-task 2, respectively. For multi-class tasks, the team scored 55 (7th) and 32 (9th) in English and Spanish, respectively.

5.12 Aruna_Devi_S

The researchers from team Aruna_Devi_S (ArunaDevi and Bharathi, 2024), they tried several ML and DL techniques as: MultinomialNB, SGDClassifier, SVM, LR and BERT model, achieving the best results with the last. No preprocessing were here specified. Consequently, they achieved an average Macro-F1 score of 41.61 (12th).

5.13 MIKHAIL

MIKHAIL team (Krasitskii et al., 2023) made some data preprocessing, as many of the rest of the teams that participated in the international shared task, removing special characters and punctuation marks and made some tokenization. They mainly used the popular transformer LLM: BERT to finally achieve the 13th position in the rank (see Table 3) with an average Macro-F1 score of 34.87.

MIKHAIL team followed the same approach to perform sub-task 2. The team scored 81 (5th) and 85(3rd) micro-F1 score in binary tasks for Spanish and English, respectively. They scored 4th place for both multi-class tasks.

5.14 Zavira

Zavira team (Ahani et al., 2024) used albert-base and mBERT models from Simple Transformers [2] library for both tasks in English and Spanish, respectively. They achieve 7th and 3rd place for binary tasks in Spanish and English, respectively. For multi-class tasks, they scored a micro-F1 score of 67(3rd) and 44 (8th) for English and Spanish.

5.15 AmnaNaseeb

AmnaNaseeb team (Naseeb et al., 2024) used traditional machine learning models: logistic regression and support vector machines, and deep neural network methods to perform subtask 2. They participated only in the English language for both binary and multi-class sub-tasks. They scored micro-F1 scores of 16(10th) and 32(10th) in binary and multi-class tasks respectively.

5.16 Lemlem

Lemlem team (Lemlem et al., 2024) used traditional machine learning models and transformer algorithms such as Support Vector Machine (SVM), Random Forest (RF), and a

[2] https://simpletransformers.ai/

transformer-based BERT model. They participated in binary and multi-class sub-task 2 tasks in English. They achieved a micro-F1 score of 73(9th) and 15(11th) in binary and multi-class classification tasks, respectively.

6 Results and discussion

6.1 Subtask 1 - EDI

The official leaderboard for subtask 1 of HOPE 2024 shared task is shown in Table 3. For subtask 1, team NTT@UIT, from University of Information Technology-VNUHCM (Vietnam) and Vietnam National University, achieved the best results, followed by ChauPhamQuocHung, CUFE and Ometeotl teams. NTT@UIT ranked with an average Macro-F1 score of 71.61 (5.82 points over the second best qualified team). NTT@UIT taeam obtained their results using ChatGPT 3.5 and they applied different prompting techniques and describing analysis estrategies but they highlighted the importance of describing the data. In the past HOPE 2023 edition, Zootopi team used ChatGPT 3.5 too to achieve the best result and, in contrast to NTT@UIT, they found specially important that the dataset were generally aligned with ChatGPT ethics, tagging as hope speech those texts that spoke in favour of the LGTB community. This team also highlighted some biased responses from ChatGPT, particularly when they were related to sensitive words, such as the word "Trump", which usually refers to the President Donald Trump (Ngo and Tran, 2023). All second, third and fourth ranked participants obtained very close average Macro-F1 results (65.79, 65.22 and 64.38, respectively) and both ChauPhamQuocHung and Ometeotl used similar techniques focused in using transformers. Nevertheless, it was not the case of CUFE team, they proposed a different aproach based on Gzip compression. In relation to the dataset, the teams did present any complaints.

6.2 Subtask 2 - PolyHope

The official results for subtask 2, Hope as Expectations, are presented in Table 4. This table evaluate the performance of participants' systems on Spanish and English datasets. The reported metrics include Precision (P), Recall (R), and F1-score (F1), both macro-averaged and weighted, as well as each team's rank based on the macro F1-score.

6.2.1 Subtask 2.a - Binary classification

Based on Table 4 for the binary subtask, in the Spanish dataset, olp emerged as the top performer with a macro F1-score of 0.85 and a weighted F1-score of 0.87, closely followed by ABCD Team with similar scores. Teams such as ChauPhamQuocHung and UMUTeam also demonstrated strong performance, securing ranks in the top three with consistent scores across the metrics. Mid-tier teams like MUCS and Ometeotl maintained competitive scores around 0.82, indicating reliable but slightly lower performance compared to the top-ranked teams. On the other hand, in the English dataset and the same table, ABCD Team led the competition with an F1-score of 0.87, while "olp" ranked second with an F1-score of 0.86, highlighting their strong and consistent performance across both languages. Other notable teams such as UMUTeam, MIKHAIL, and Zavira showed robust results, achieving F1-scores around 0.85, indicating their effectiveness in binary classification tasks. The performance of lower-tier teams such as ngocthien705b, Girma and NLP_URJC was moderate, with F1-scores ranging from 0.76 to 0.79, while teams like NTT@UIT and CIC-PAK scored lower, around 0.73 to 0.76. The weakest performers, AmnaNaseeb and Juan-Calderon, had significantly lower F1 scores, reflecting challenges in achieving high accuracy and reliability in their models.

The analysis highlights the competitive nature of the Binary setting for subtask 2, with the olp and ABCD Team leading in both Spanish and English languages. The metrics suggest a high level of consistency among the top teams, particularly in achieving balanced Precision and Recall scores, leading to strong F1 scores. The detailed breakdown of macro and weighted scores provides insights into the performance stability across different evaluation measures, emphasizing the effectiveness of the top-performing systems.

6.2.2 Subtask 2.b - Multiclass classification

Table 4 present the performance of different teams in Subtask 2.b Multiclass PolyHope, comparing results across Spanish and English languages.

In Spanish, the ABCD Team leads with a macro F1 score of 0.67 and a weighted F1

			subtask 1.a			subtask 1.b		
		Avg.	Macro scores			Macro scores		
#	Team	Macro-F1	P	R	F1	P	R	F1
01	NTT@UIT	71.61	65.22	90.00	75.63 (01)	62.18	74.00	67.58 (01)
02	ChauPhamQuocHung	65.79	74.00	74.00	74.00 (02)	58.16	57.00	57.58 (04)
03	CUFE	65.22	64.15	68.00	66.02 (04)	62.04	67.00	64.42 (02)
04	Ometeotl	64.38	64.22	70.00	66.99 (03)	64.84	59.00	61.78 (03)
05	ABCD Team	58.79	80.95	51.00	62.58 (06)	73.33	44.00	55.00 (05)
06	MUCS	52.31	69.51	57.00	62.64 (05)	54.84	34.00	41.98 (08)
07	VEL	51.03	76.27	45.00	56.60 (08)	64.81	35.00	45.45 (07)
08	CICPAK	49.48	62.50	45.00	52.33 (09)	60.32	38.00	46.63 (06)
09	UMUTeam	48.90	88.68	47.00	61.44 (07)	60.47	26.00	36.36 (11)
10	UniOfGalway	46.66	64.18	43.00	51.50 (10)	60.38	32.00	41.83 (09)
11	Girma	44.40	66.10	39.00	49.06 (11)	58.82	30.00	39.74 (10)
12	Aruna_Devi_S	41.61	76.60	36.00	48.98 (12)	54.35	25.00	34.25 (12)
13	MIKHAIL	34.87	80.65	25.00	38.17 (13)	63.64	21.00	31.58 (13)

Table 3: HOPE 2024@IberLEF subtask 1 results. Teams are sorted by their average Macro-F1 score.

score of 0.82, followed closely by the team olp and ChauPhamQuocHung with macro F1 scores of 0.66 and 0.65, respectively, both maintaining a weighted F1 score of 0.81 or 0.80. The rankings show a gradual decline in scores among the teams, with CICPAK having the lowest macro F1 score of 0.30 and a weighted F1 score of 0.59.

For English, the ABCD Team also leads with a macro F1 score of 0.72 and a weighted F1 score of 0.78, sharing the top rank with olp and ChauPhamQuocHung, who have similar macro and weighted scores. Teams like MIKHAIL and Zavira have notable performances as well, but a steep decline is observed, with AmnaNaseeb and Lemlem scoring the lowest macro F1 scores of 0.16 and 0.15, respectively, indicating significant room for improvement. Overall, the performance across both languages highlights the varying strengths and areas for development among the participating teams.

The results highlight that the ABCD, olp and ChauPhamQuocHung teams consistently achieved the highest performance across both Spanish and English tasks, while other teams showed varying degrees of proficiency, indicating a need for improvement among lower-ranked teams.

7 Conclusions and future work

This paper presents the description of the second international shared task on multilingual hope speech detection, organized within the IberLEF workshop, in the framework of the SEPLN 2024 conference. Specifically, we proposed two subtasks, the detection of hope speech for EDI texts and the identification of hope speech focusing on expectations, and desirable and undesirable facts. Thirteen different teams participated in the first subtask and seventeen in the second subtask, with a total of nineteen different teams. The best result from the EDI subtask (NTT@UIT) achieved an average macro F1-score of 71.61 while in the second task, which targeted the identification of hope speech as expectations, the ABCD Team led the competition across both binary and multiclass classifications in both Spanish and English datasets. Specifically, in the binary classification task, ABCD Team achieved an outstanding micro F1-score of 87 for English and 86 for Spanish, underscoring their robust performance in accurately distinguishing between hope and non-hope speech. In the multiclass classification task, ABCD Team also demonstrated strong performance with a macro F1-score of 0.78 in English and 0.82 in Spanish, reaffirming their expertise across different classification challenges.

In future work, we plan to improve the datasets by increasing their sizes to further promote the detection of hope speech. Specifically, in relation to the first subtask, we plan to increase the number of examples for differ-

Languages	**Spanish/English**						
Team	**Macro scores**			**Weighted scores**			**Rank**
	P	**R**	**F1**	**P**	**R**	**F1**	**M_F1**
Subtask 2.a							
olp	0.85/0.86	0.85/0.86	0.85/0.86	0.87/0.86	0.87/0.86	0.87/0.86	1/2
ABCD Team	0.84/0.87	0.85/0.87	0.84/0.87	0.86/0.87	0.86/0.87	0.86/0.87	2/1
ChauPham	0.84/0.85	0.82/0.85	0.83/0.85	0.85/0.85	0.86/0.85	0.85/0.85	3/3
UMUTeam	0.83/0.86	0.83/0.86	0.83/0.86	0.85/0.86	0.85/0.86	0.85/0.86	3/2
MUCS	0.82/0.82	0.82/0.82	0.82/0.82	0.84/0.82	0.84/0.82	0.84/0.82	4/5
Ometeotl	0.82/0.82	0.81/0.82	0.81/0.82	0.84/0.82	0.84/0.82	0.84/0.82	5/5
MIKHAIL	0.80/0.85	0.81/0.84	0.81/0.85	0.83/0.85	0.83/0.85	0.83/0.85	5/3
hamadanayel	0.81/0.83	0.79/0.83	0.80/0.83	0.82/0.83	0.83/0.83	0.82/0.83	6/4
Zavira	0.82/0.85	0.77/0.85	0.79/0.85	0.82/0.85	0.83/0.85	0.82/0.85	7/3
NLP_URJC	0.80/0.79	0.76/0.79	0.77/0.79	0.81/0.79	0.81/0.79	0.81/0.79	8/7
ngocthien705b	0.81/0.81	0.74/0.81	0.76/0.81	0.81/0.81	0.81/0.81	0.80/0.81	9/6
Girma	0.80/0.82	0.73/0.82	0.75/0.82	0.80/0.82	0.80/0.82	0.79/0.82	10/5
NTT@UIT	0.73/0.75	0.73/0.74	0.73/0.73	0.76/0.75	0.76/0.73	0.76/0.73	11/9
CICPAK	0.71/0.74	0.72/0.74	0.71/0.74	0.75/0.74	0.74/0.74	0.74/0.74	12/8
Lemlem	-/0.73	-/0.73	-/0.73	-/0.74	-/0.73	-/0.73	-/9
AmnaNaseeb	-/0.24	-/0.50	-/0.32	-/0.23	-/0.48	-/0.31	-/10
JuanCalderon	-/0.20	-/0.20	-/0.20	-/0.21	-/0.21	-/0.21	-/11
Subtask 2.b							
ABCD Team	0.68/0.71	0.65/0.73	0.67/0.72	0.82/0.78	0.82/0.77	0.82/0.78	1/1
olp	0.66/0.72	0.67/0.72	0.66/0.72	0.82/0.77	0.81/0.77	0.81/0.77	2/1
ChauPham	0.65/0.73	0.65/0.72	0.65/0.72	0.81/0.78	0.80/0.78	0.80/0.78	3/1
MIKHAIL	0.64/0.64	0.65/0.67	0.64/0.65	0.79/0.73	0.79/0.72	0.79/0.72	4/4
MUCS	0.63/0.59	0.65/0.54	0.64/0.59	0.80/0.67	0.79/0.68	0.80/0.67	4/6
UMUTeam	0.65/0.69	0.61/0.68	0.63/0.68	0.79/0.76	0.79/0.75	0.79/0.75	5/2
hamadanayel	0.63/0.63	0.47/0.57	0.50/0.59	0.73/0.69	0.76/0.68	0.73/0.69	6/5
NTT@UIT	0.47/0.46	0.51/0.43	0.48/0.42	0.70/0.56	0.67/0.62	0.68/0.58	7/9
Zavira	0.51/0.67	0.44/0.68	0.44/0.67	0.67/0.74	0.69/0.74	0.67/0.74	8/3
Girma	0.49/0.59	0.39/0.53	0.32/0.55	0.62/0.64	0.68/0.66	0.61/0.65	9/3
CICPAK	0.47/0.54	0.30/0.44	0.30/0.45	0.60/0.58	0.66/0.59	0.59/0.56	10/8
AmnaNaseeb	-/0.12	-/0.25	-/0.16	-/0.23	-/0.48	-/0.31	-/10
Lemlem	-/0.14	-/0.16	-/0.15	-/0.32	-/0.38	-/0.35	-/11

Table 4: HOPE 2024@IberLEF Subtask 2.a and Subtask 2.b Binary PolyHope Spanish/English results.

ent topics from LGBT in SpanishHopeMultidomain to make it a more capable multidomain automatic hope speech detection dataset. For Hope as expectations, we plan to extend the dataset in different dimensions, including low-resource languages and culturally aware study of expectations.

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, and Project FedDAP (PID2020-116118GA-I00) and Project Trust-ReDaS (PID2020-119478GB-I00) supported by MICINN/AEI/10.13039/501100011033.
It is also part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe, LT-SWM (TED2021-131167B-I00) funded

by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR and "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. The research work conducted by Salud María Jiménez-Zafra has been supported by Action 7 from Universidad de Jaén under the Operational Plan for Research Support 2023-2024.

References

Ahani, Z., M. S. Tash, M. Tash, A. Gelbukh, and I. Gelbukh. 2024. Multiclass hope speech detection through transformer methods. 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*.

Armenta-Segura, J. and G. Sidorov. 2024. Ometeotl at hope2024@iberlef: Custom bert models for hope speech detection. 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*.

ArunaDevi, S. and B. Bharathi. 2024. Machine learning based approach for hope speech detection. 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*.

Bade, G. Y., O. Koleniskova, J. L. Oropeza, G. Sidorov, and K. F. Bergene. 2024. Hope speech in social media texts using transformer. 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*.

Balouchzahi, F., G. Sidorov, and A. Gelbukh. 2023. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Bruininks, P. and B. F. Malle. 2005. Distinguishing hope from optimism and related affective states. *Motivation and Emotion*, 29:324–352.

Bui Hong, S., Q. Le Minh, and T. Dang Van. 2024. ABCD Team at HOPE 2024: Hope Detection with BERTology Models and Data Augmentation. 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*.

Burnap, P., G. Colombo, R. Amery, A. Hodorog, and J. Scourfield. 2017. Multi-class machine classification of suicide-related communication on twitter. *Online Social Networks and Media*, 2:32–44, 08.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Chakravarthi, B. R. 2020. 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*, pages 41–53, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Chakravarthi, B. R., 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, and J. García-Díaz. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 378–388, Dublin, Ireland, May. Association for Computational Linguistics.

Chau Pham Quoc, H. and T. Dang Van. 2024. Choosing the Right Language Model for the Right Task. 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.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. 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.*

Divakaran, S., K. Girish, and H. L. Shashirekha. 2024. Hope on the horizon: Experiments with learning models for hope speech detection in spanish and english. 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.*

Eid, M. and R. Larsen. 2008. *The Science of Subjective Well-Being*. Guilford Publications.

Ibrahim, M. 2024. Parameter-free spanish hope speech detection. 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.*

Jiménez-Zafra, S. M., M. Á. García, D. García-Baena, J. García-Díaz, B. R. Chakravarthi, R. Valencia-García, and L. A. Ureña-López. 2023. Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection. *Procesamiento del Lenguaje Natural*, 71(0):371–381.

Kitzie, V. 2018. "i pretended to be a boy on the internet": Navigating affordances and constraints of social networking sites and search engines for lgbtq+ identity work. *First Monday*, 23, 07.

Krasitskii, M., O. Kolesnikova, L. C. Hernandez, G. Sidorov, and A. Gelbukh. 2023. Hope2023@iberlef: A cross-linguistic exploration of hope speech detection in social media. 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.*

Kumaresan, P. K., B. R. Chakravarthi, S. Cn, M. Á. García-Cumbreras, S. M. Jiménez Zafra, J. A. García-Díaz, R. Valencia-García, M. Hardalov, I. Koychev, P. Nakov, D. García-Baena, and K. K. Ponnusamy. 2023. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In B. R. Chakravarthi, B. Bharathi, J. Griffith, K. Bali, and P. Buitelaar, editors, *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–53, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Lemlem, E., Y. Tsadkan, N. Amna, S. Grigori, and B. Ildar. 2024. Enhancing hope speech detection on twitter using machine learning and transformer models. 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.*

Milne, D., G. Pink, B. Hachey, and R. Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. pages 118–127, 01.

Naseeb, A., K. L. Eyob, G. Sidorov, and O. Kolesnikova. 2024. Hope@iberlef 2024: Beyond binary bounds—classifying hope in online discourse. 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.*

Ngo, A. and H. T. H. Tran. 2023. Zootopi at hope2023iberlef: Is zero-shot chatgpt 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), CEUR-WS.org.*

Nguyen Thi, T. and T. Dang Van. 2024. An Empirical Study of Prompt Engineering with Large Language Models for Hope

Detection in English and Spanish. 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*.

Palakodety, S., A. R. KhudaBukhsh, and J. G. Carbonell. 2019. Hope speech detection: A computational analysis of the voice of peace. *arXiv preprint arXiv:1909.12940*.

Pan, R., Ángela Almela, and G. Alcaraz-Mármol. 2024. UMUTeam at hope@iberlef 2024: Fine-tuning approach with sentiment and emotion features for hope speech detection. 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*.

Pérez, J. M., D. A. Furman, L. Alonso Alemany, and F. M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June. European Language Resources Association.

Pérez, J. M., M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. Vanina Martínez. 2023. pysentimiento: A python toolkit for opinion mining and social nlp tasks.

Ponnusamy, K. K., M. Vegupatti, P. K. Kumaresan, R. Priyadharshini, P. Buitelaar, and B. R. Chakravarthi. 2024. VEL@iberlef 2024: Hope speech detection in spanish social media comments using bert pre-trained model. 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*.

Snyder, C., B. Hoza, W. Pelham, M. Rapoff, L. Ware, M. Danovsky, L. Highberger, H. Ribinstein, and K. Stahl. 1997. The development and validation of the children's hope scale. *Journal of Pediatric Psychology - J PEDIAT PSYCHOL*, 22:399–421, 06.

Snyder, C. 2002. Hope theory: Rainbows in the mind. *Psychological Inquiry*, 13:249–275, 10.

Snyder, C. 1994. *The Psychology of Hope: You Can Get There from Here*. Free Press.

Subburaj, A., A. Kathiresan, R. Ponnusamy, P. Butilaar, and B. R. Chakravarthi. 2024. UniOfGalway@iberlef 2024: Hope speech recognition in spanish: A comparative analysis of transformer-based models. 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*.

Ullah, F., M. T. Zamir, M. Ahmad, G. Sidorov, and A. Gelbukh. 2024. Hope: A multilingual approach to identifying positive communication in social media. 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*.