Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 371-381 recibido 24-06-2023 revisado 09-07-2023 aceptado 19-07-2023

# Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection

## *Resumen de la tarea HOPE en IberLEF 2023: Detección Multilingüe de Discurso Esperanzador*

**Salud María Jiménez-Zafra,**[1] **Miguel Ángel García-Cumbreras,**[1] **Daniel García-Baena,**[1]
**José Antonio García-Díaz,**[2] **Bharathi Raja Chakravarthi,**[3]
**Rafael Valencia-García,**[2] **L. Alfonso Ureña-López**[1]
[1]Computer Science Department, SINAI, CEATIC, Universidad de Jaén
[2]Facultad de Informática, Universidad de Murcia
[3]School of Computer Science, University of Galway
{sjzafra,magc,laurena}@ujaen.es, daniel.gbaena@gmail.com
{joseantonio.garcia8,valencia}@um.es
bharathi.raja@universityofgalway.ie

**Abstract:** Hope speech is the speech that is able to relax hostile environments and that helps, inspires and encourages people in times of illness, stress, loneliness or depression. Its automatic recognition can have a very significant effect fighting against sexual and racial discrimination or fostering less belligerent environments. In contrast to identifying and censoring negative or hate speech, hope speech detection is focused on recognizing and promoting positive speech online. In this paper we present an overview of the IberLEF 2023 shared task, HOPE: Multilingual Hope Speech Detection, consisting of identifying whether texts written in English or Spanish contain hope speech or not. The competition was organized through CodaLab and attracted 50 teams that registered. Finally, 12 submitted results and 8 presented working notes describing their systems.
**Keywords:** Hope speech detection, natural language processing, equality, diversity and inclusion.

**Resumen:** Definimos el discurso de la esperanza como aquel que es capaz de relajar entornos hostiles y que ayuda, inspira y anima a las personas en momentos de enfermedad, estrés, soledad o depresión. Su detección automática puede tener un efecto muy significativo luchando contra la discriminación sexual y racial o fomentando entornos menos beligerantes. A diferencia de la identificación y censura del discurso negativo o de odio, la detección del discurso esperanzador se centra en reconocer y promover el discurso positivo. En este artículo presentamos los resultados de la tarea de IberLEF 2023, HOPE: Detección multilingüe del discurso de la esperanza, que consiste en identificar si textos escritos en inglés o español contienen o no discurso de esperanza. La competición se organizó a través de CodaLab y atrajo a 50 equipos que se inscribieron. Finalmente, 12 equipos presentaron resultados y 8 enviaron artículos describiendo sus sistemas.
**Palabras clave:** Detección de discurso esperanzador, procesamiento de lenguaje natural, igualdad, diversidad e inclusión.

### *1 Introduction*

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). Detect it automatically, so that positive comments can be more widely disseminated, have a significant effect to combating sexual or racial discrimination, and to foster a less belligerent environment. (Palakodety, KhudaBukhsh, and Carbonell, 2019).

On social media, offensive messages are posted against people because of their race, color, ethnicity, gender, sexual orientation, nationality, or religion. As Chakravarthi

ISSN 1135-5948 DOI 10.26342/2023-71-29 

(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.

The *HOPE: Multilingual Hope Speech Detection* shared task is related to the inclusion of vulnerable groups and focuses on the study of the detection of hope speech, in pursuit of Equality, Diversity and Inclusion (EDI). It consists of, given a text, written in Spanish or English, identifying whether it contains hope speech or not. This shared task is organized at IberLEF 2023 (Jiménez-Zafra, Rangel, and Montes-y Gómez, 2023), as part of the XXXIX International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023).

This task was previously organized at the second workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI-2022), as part of ACL 2022, but for five languages: Tamil, Malayalam, Kannada, English and Spanish (Chakravarthi et al., 2022). The novelties of this shared task are threefold: i) it is organized in two languages, Spanish and English; ii) it provides an expanded and improved dataset; and iii) it is directed to the IberLEF community.

## *2 Task description*

The aim of the HOPE 2023 shared task is to promote EDI through the detection of hope speech. It consists of identifying whether a text, written in Spanish or English, contains hope speech or not.

The general challenges proposed for this first edition are:

1. To promote research in inclusive Language Technologies (LT).
2. To adopt and adapt appropriate LT models to suit hope speech.
3. To provide opportunities for researchers from the LT community around the world to collaborate with other researchers to identify and propose possible solutions for the challenges of hope speech.

Some specific challenges of the task are the following:

1. Identifying hope speech in two languages: Spanish and English.
2. Dealing with two different social networks: Twitter and YouTube.
3. Lack of context: Tweets are short (up to 240 characters) and they are annotated separately without data about the tweets they are reply to, as the annotators do not have access to the rest of the tweets in the conversation.
4. Informal language: misspellings, emojis and onomatopoeias are common.

This shared task is divided into two subtasks, according to the language in which the texts are written. These subtasks are described below.

### 2.1 Subtask 1: Hope speech detection in Spanish

This subtask consists of, given a Spanish tweet, identifying whether it contains hope speech or not. The possible categories for each text are:

- HS: Hope Speech.
- NHS: Non-Hope Speech.

### 2.2 Subtask 2: Hope speech detection in English

This subtask consists of, given an English YouTube comment, identifying whether it contains hope speech or not. The possible categories for each text are:

- HS: Hope Speech.
- NHS: Non-Hope Speech.

The competition was organized through CodaLab and is available at the following link: `https://codalab.lisn.upsaclay.fr/competitions/10215`. In both subtasks there was a real time leaderboard and the participants were allowed to make a maximum of 10 submissions through CodaLab for each subtask, from which each team had to select the best ones for ranking. The participant systems were evaluated using precision, recall and F1-score per category and averaged using the macro-average method. Finally, they were ranked using the macro-average F1-score.

## 3 Datasets

We provided two datasets, one with texts written in Spanish for subtask 1 and the other in English for subtask 2.

The Spanish dataset is an improved and extended version of the SpanishHopeEDI dataset (García-Baena et al., 2023). The SpanishHopeEDI dataset was improved by manual revision of the annotations, as some annotation errors were found in the error analysis of the baseline experiments conducted with the dataset (García-Baena et al., 2023). It consists of LGTB-related tweets that were collected with the Twitter API (June 27, 2021 to July26, 2021) and using a lexicon of LGBT-related terms, such as #OrgulloLGTBI or #LGTB, as seed for the search. This dataset was extended with a set of tweets collected using the UMUCorpusClassifier tool (García-Díaz et al., 2020), which allows defining different search criteria such as keywords, accounts and geolocation. The keywords used to collect the tweets were related to transphobia and homophobia, such as #transfobia (*#transphobia*), transexual (*transsexual*), identidad de género (*gender identity*), #homofobia (*#homophobia*), homosexual (*homosexual*), #AlertaHomofobia (*#HomophobiaAlert*), or #StopLGTBIfobia *#StopLGTBIphobia*. It should be mentioned that all the tweets of this dataset were manually labelled by the organizers of the shared task marking a tweet as HS (hope speech) if the text: i) explicitly supports the social integration of minorities; ii) is a positive inspiration for the LGTB community; iii) explicitly encourages LGTB people who might find themselves in a situation; or iv) unconditionally promotes tolerance. On the contrary, a tweet was marked as NHS (non-hope speech) if the text: i) expresses negative sentiment towards the LGTB community; ii) explicitly seeks violence; or iii) uses gender-based insults.

On the other hand, the English dataset was harvested from YouTube, and it is compound by comments that were posted into that video streaming platform. The subject matter of the comments written in English were *EDI*, including women in the *STEM* group and people from the *LGBT* collective, COVID-19, the *Black Lives Matters* movement, UK versus China, US versus China, and Australia versus China (Chakravarthi, 2020; Chakravarthi et al., 2022). This dataset was manually annotated by external annotators using the guidelines defined in the work of Chakravarthi (2020).

Table 1 and Table 2 show the distribution of both datasets considering the number of samples for each label and set.

| Type | Dev | Train | Test |
|---|---|---|---|
| Hope | 100 | 691 | 150 |
| Non-Hope | 200 | 621 | 300 |
| **Total** | **300** | **1,312** | **450** |

Table 1: Spanish dataset statistics.

| Type | Dev | Train | Test |
|---|---|---|---|
| Hope | 268 | 1,961 | 21 |
| Non-Hope | 2,531 | 20,690 | 4,784 |
| **Total** | **2,799** | **22,651** | **4,805** |

Table 2: English dataset statistics.

## 4 Task Settings

This shared-task was organized through CodaLab and is 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 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, 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. The systems were evaluated using precision, recall and F1-score per category and averaged using the macro-average method.

### 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 2023 shared task attracted 50 teams that registered through CodaLab, of which 12 submitted results and 8 presented working notes describing their systems. In this section, we briefly describe each of the proposals submitted by the participants.

### 5.1 Habesha

The members of the Habesha team (Gemeda Yigezu et al., 2023) (Instituto Politécnico Nacional, Centro de Investigación en Computación, México) developed a model based on Support Vector Machines (SVM). They used term frequency and inverse document frequency (TF–IDF) for feature selection and they obtained an average macro F1 for English of 0.489 and 0.481 for Spanish.

The researchers considered that the model performed better in English than in Spanish because the Spanish dataset is significantly smaller than the English one. This team ranked 10th in the Spanish subtask. In the English subtask, they forgot to publish their results in the official leaderboard. If they had published them, they would have achieved the 6th position in English.

### 5.2 I2C-Huelva

In relation to the first subtask, the I2C-Huelva team (Domínguez Olmedo, Mata Vázquez, and Pachón Álvarez, 2023) (I2C Research Group, University of Huelva, Spain) employed BERTuit (Huertas-Tato, Martin, and Camacho, 2022), a transformer proposed for Spanish language, pre-trained using RoBERTa (Liu et al., 2019) optimization. BERTuit was trained from scratch with texts created by native speakers and published on Twitter from 2021 to 2018. This team applied a basic preprocessing consisted of:

1. Changing texts to lowercase.
2. Removing http/https links.
3. Removing all hash signs (#).
4. Strip white spaces (including newlines).

For the second subtask, hope speech detection in English, they employed DistilBERT model (Sanh et al., 2019). In addition, a basic preprocessing was applied to all texts:

1. Changing text to lowercase.
2. Removing http/https links.
3. Removing the hash signs (#).
4. Removing the @ signs.
5. Removing the emojis.
6. Striping white spaces (including newlines).
7. Removing Unicode characters.
8. Removing single letters and numbers surrounded by spaces.

With both models, BERTuit and DistilBERT, they used the default hyperparameters shown below:

- Optimizer: AdamW.
- Radam_epsilon: 1e-8.
- learning_rate: 4e-5.
- Strip white spaces (including newlines).
- train_batch_size: 8.

This team achieved the 2nd best results for the Spanish subtask, with an average macro F1 value of 0.744, and the 1st position for the English subtask (average macro F1 value of 0.501). They tried several configurations of BERTuit (subtask 1) and DistilBERT (subtask 2) and their best results were obtained when they preprocesed the texts.

### 5.3 LIDOMA

The researchers from LIDOMA team (Shahiki-Tash et al., 2023) (Instituto Politécnico Nacional, Centro de Investigación en Computación, México) used a 5-layered Convolutional Neural Networks (CNNs) to embed the samples from a Bag of Words (BoW) representation and retrieved

relevant lexical features. They firstly preprocessed all texts removing emoticons, special characters, pictographs, flags, transport and map symbols, as well as URL patterns. Additionally, this team lower-cased all texts.

The final step was to train their 5-layered CNN using Keras. In the first layer, an upper bound on the accepted features needed to be specified, which was then embedded in the second layer with a fixed dimension. Next, in the convolutional layer, the size and number of kernels were specified, as well as the activation function. The team chose ReLU as the optimal activation function. For the output layer, the sigmoid function was chosen. Finally, an L2 penalty on the kernels was added to prevent overfitting.

In the Spanish dataset, their recall was 0.7467 for HS and 0.853 for NHS. Precision was 0.586 for HS and 0.853 for NHS. Authors believe that the low precision in HS was due to several hope speech samples that, although supportive of LGBT+ issues, also included violent language or offensive slurs.

The English dataset, on the other hand, was extremely unbalanced, as it can be seen in Table 2, with only 8.76% HS samples and 91.24% NHS samples. The authors considered that the dataset imbalance has led to obtain a generally high precision in NHS samples but a very low precision in HS samples.

This team ranked in 4th position in both subtasks.

### 5.4 NLP_SSN_CSE

The NLP_SSN_CSE team (Balaji et al., 2023) (Department of CSE, Sri SivaSubramaniya Nadar College of Engineering, Tamil Nadu, India) preprocessed all texts removing the HTML tags, hashtags, social media mentions, and URLs. Emojis and emoticons were replaced with the text they stand for. In addition, short terms were extended and everything was lowercased. Finally, any extraneous white spaces were eliminated too.

In relation to feature extraction, CountVectorizer and TF–IDF vectorizer were used. Researchers classified data using several deep learning and traditional machine learning models such as multilingual BERT (mBERT) (Devlin et al., 2019), BERT (Devlin et al., 2019), Random Forest, SVM, Logistic Regression, and Decision Tree. It is worth nothing that all transformers used the uncased version.

Among all the models trained, mBERT provided the best results for the English and Spanish datasets with a weighted F1-score of 92.87% and 96.57% respectively, for the evaluation dataset. The Logistic Regression and Random Forest classifiers achieved similar F1-scores of 92.10% and 92.07% for the English dataset using TF–IDF vectorizer. However the results obtained for Spanish were comparatively low.

This team obtained the 8th position for the Spanish subtask and the 5th position for the English one.

### 5.5 NLP_URJC

The researchers from NLP_URJC team (Rodríguez-García, Riaño Martínez, and Montalvo Herranz, 2023) (Universidad Rey Juan Carlos, Spain) employed BERT (Devlin et al., 2019) in the English subtask, and the Spanish version of BERT, BETO (Cañete et al., 2020) in the Spanish subtask. Both models were integrated into a pipeline to address the subtasks demanded in the challenge. This pipeline consisted of four steps: preprocessing, building the selected model, training and testing.

The pipeline worked as follows: first, the input data was processed in the cleaning phase where URLs, emojis and stop words were removed. Second, a model was built and fine-tuned depending on the subtask addressed. Finally, the assessment phase was conducted, and the development dataset was employed to assess its accuracy. As a result, three different outcomes could be obtained: a file with the samples classified, a picture with the resulting confusion matrix and the evaluation metrics values.

This team highlights the difference between the good results obtained working with the Spanish corpus versus the disappointing ones from the English dataset. Researchers believe that this behavior is due to an unusual distribution of the tweets in the datasets. They used a word cloud with the English dataset to show that there are no highly discriminating features which may help the model discern between the two labels, finally selecting the NHS category since it is the majority class.

They would ranked 8th (0.657) for the Spanish subtask and 1st for the English one (0.502), but they submitted their results out of time.

### 5.6 UMUTeam

The members of the UMUTeam (Pan, Alcaraz-Mármol, and García-Sánchez, 2023) (Universidad de Murcia, Spain) preprocessed the datasets by replacing all of the hashtags and mentions with #[HASHTAG] and @[USER], and changing all the emojis by their textual meaning, using the emoji library. In addition, for English, general contractions were expanded through the contraction library. Then, authors employed data augmentation techniques to enhance the overall performance of their classification models, translating the texts classified as hope speech from the English dataset into Spanish, for subtask 1, and the other way around for subtask 2.

In relation to the classification model, they made use of finetuning with some different pretrained models based on the transformers architecture to develop their final classification models for both subtask 1 and subtask 2. For the classification of the hope speech texts, the researchers incorporated an additional sequence classification layer, at the end of the LLMs, which allowed the system to output better classification results. The LLMs that the team employed were: BETO (Cañete et al., 2020), ALBETO (Cañete et al., 2022), DistilBETO (Cañete et al., 2022), MarIA (no et al., 2022), XLM-R (Conneau et al., 2020), BERT (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019) and DeBERTa-large (He et al., 2020).

All models were finetuned using the hyperparameters from below:

- Training batch size of 16, 15 epochs.
- Learning rate: 2e-5.
- Weight decay: 0.01

In the Spanish subtask, they ranked in 5th position, with a macro F1-score of 0.710, while in the English subtask they were placed in 7th position, with a macro F1-score of 0.482.

### 5.7 Zavira

The team Zavira (Ahani et al., 2023) (Instituto Politécnico Nacional, Centro de Investigación en Computación, México) used the SVM algorithm for the English dataset and the K-Nearest Neighbors (KNN) algorithm for the Spanish dataset.

The authors extensively preprocessed all data. They used a lemmatizer to reduce words to their base form and the removal of stop words and punctuation marks to simplify the text. Additionally, they employed a clean-text method to remove any irrelevant or redundant information that might negatively impact the performance of the models. To generate features, the researchers employed the Scikit-learn module's TF–IDF Vectorizer to extract character n-grams from pre-processed textual data that had already been cleaned and lemmatized.

Researchers suggested that cross-lingual classification of hope and non-hope texts is a challenging task, particularly due to the linguistic differences between languages.

The team ranked in the 3rd position in both subtasks. Specifically, the SVM algorithm obtained an F1-score of 0.490, while the KNN-based approach got an F1-score of 0.740.

### 5.8 Zootopi

The Zootopi team (Ngo and Tran, 2023) (Wroclaw University of Science and Technology, Poland; Jožef Stefan International Postgraduate School, Slovenia; Jožef Stefan Institute, Slovenia and University of La Rochelle, France) considered the two substasks as a sequence classification task. They examined if the language model improved its performance when additional training data from another language was added (data augmentation). They also checked the capacity of XLM-R (Conneau et al., 2020) model in order to incorporate knowledge working with some language different from the one used during the pretraining phase. In addition, they evaluated ChatGPT's performance and examined how significantly it was affected by certain sensitive words, which could potentially lead to data poisoning.

More specifically, they tested XLM-R with some different monolingual, crosslingual and multilingual setups. On the other hand, in relation to ChatGPT, researchers formulated two different prompting scenarios respectively for the English and Spanish datasets. They found interesting to test this two different approaches because the Spanish dataset focuses solely on the LGBT+ community, and therefore, the prompt furnished both sentence context and criteria for determining whether a text is classified as HS or

NHS. Nevertheless, regarding to the English dataset, as context were related to different fields, they developed a second scenario were they provided another different domain-specific information to the chatbot.

Regarding to the sequence-classification mechanism, the results demonstrated that rich-resourced cross-lingual learning can significantly enhance the model's performance. Specifically, the cross-lingual setup in Spanish outperforms the monolingual setup by up to 7 percentage points (pp) in F1-score and up to 5 pp compared to the multilingual setup. However, for rich-resourced languages, the monolingual setup performs better without the need for additional knowledge from other less-resourced languages.

On the other hand, the best prompting with ChatGPT surpassed the performance of other sequence-classification approaches the authors applied with a large margin for all three evaluation metrics regarding the clean and less-resourced language, Spanish, but failed to capture the information from English contexts to provide a good classification.

Finally, the researchers highlighted some biased responses from ChatGPT, particularly when it relates to sensitive words, such as the word "Trump", which usually refers to President Donald Trump. They also commented that as most of the instances in the Spanish dataset clearly convey an attitude or sentiment towards the LGBT+ community. They think that this also explains why ChatGPT performs significantly better than other methods as well as it does on the English dataset.

This team ranked in 1st position in the Spanish subtask and 9th in the English subtask.

## *6 Results and discussion*

The official leaderboards for subtask 1 and subtask 2 of HOPE 2023 shared task are shown in Table 3 and Table 4, respectively.

For the Spanish subtask, Zootopi achieved the best results, followed by I2C-Huelva and Zavira teams. Zootopi ranked with an average macro F1 of 0.916, (0.1724 points over the second qualified). Zootopi obtained their best results using ChatGPT and they suggested that it was very important for their system that the Spanish 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.

For the English subtask, the results show less variance than for the Spanish subtask. The difference between the first (I2C-Huelva, macro F1 of 0.5012) and the last classified (Zootopi, macro F1 of 0.4429) is only 0.0583. We assume that the main reason for these results is that only 21 documents were labelled as Hope compared to the 4,784 documents labelled as Non-Hope. It is worth noting that the training split for the English dataset has the same ratio between training and testing.

In both subtasks, transformer-based models have been used, specially at the top of the leaderboard. In the Spanish subtask, Zootopi, I2C-Huelva and UMUTeam, three of the top five teams, developed a transformer-based solution. The majority of the results were very similar. For instance, in the Spanish subtask, the results between the 2nd position and 6th position differ only in 4.03 points. The exception is the 1st position that outperforms the 2nd position with 17.24 points. In case of the English subtask, as commented, the results do not have much variance. As not all the teams used transformer based solutions, we cannot conclude that transformers were finally superior than other alternatives, like classic machine learning techniques as KNN (Zavira team) or CNN (LIDOMA team), nor newer solutions as ChatGPT (Zootopi team). On the other hand, we found that teams that extensively preprocessed the data, as is the case of I2C-Huelva or Zavira, obtained substantially better results.

In relation to the datasets, some teams considered that the Spanish one was not large enough and they related their not so promising results to this. The English dataset was the most criticized between the participants because of its uneven distribution of hope speech and non-hope speech texts. Certainly, all the average macro F1 results from the English subtask were close to 0.5, so we definitely should consider some problems within the English dataset that would probably be related with its not well balanced distribution.

In any case, results for the Spanish sub-

| | | | HS | | | NHS | | |
|---|---|---|---|---|---|---|---|---|
| # | **Team** | **Score** | P | R | F1 | P | R | F1 |
| 01 | Zootopi | 91.61 | 86.71 (02) | 91.33 (01) | 88.96 (01) | 95.55 (01) | 93.00 (06) | 94.26 (01) |
| 02 | I2C-Huelva | 74.37 | 90.91 (01) | 46.67 (04) | 61.67 (05) | 78.55 (05) | 97.67 (02) | 87.07 (02) |
| 03 | Zavira | 74.30 | 62.15 (07) | 73.33 (03) | 67.28 (02) | 85.35 (02) | 77.67 (08) | 81.33 (06) |
| 04 | LIDOMA | 72.38 | 58.64 (08) | 74.67 (02) | 65.69 (03) | 85.33 (03) | 73.67 (09) | 79.07 (08) |
| 05 | UMUTeam | 71.03 | 56.99 (09) | 73.33 (03) | 64.14 (04) | 84.44 (04) | 72.33 (10) | 77.92 (10) |
| 06 | honghanhh | 70.34 | 77.65 (05) | 44.00 (05) | 56.17 (06) | 76.99 (06) | 93.67 (05) | 84.51 (03) |
| 07 | juanmanuel.calvo | 66.26 | 83.61 (03) | 34.00 (06) | 48.34 (07) | 74.55 (07) | 96.67 (03) | 84.18 (04) |
| 08 | NLP_SSN_CSE | 59.13 | 82.93 (04) | 22.67 (07) | 35.60 (08) | 71.64 (08) | 97.67 (02) | 82.65 (05) |
| 09 | aswathyprem | 48.64 | 73.68 (06) | 9.33 (09) | 16.57 (10) | 68.45 (09) | 98.33 (01) | 80.71 (07) |
| 10 | Habesha | 48.15 | 33.33 (10) | 16.67 (08) | 22.22 (09) | 66.67 (10) | 83.33 (07) | 74.07 (11) |
| 11 | mgraffg | 41.98 | 25.00 (11) | 3.33 (10) | 5.88 (11) | 66.28 (11) | 95.00 (04) | 78.08 (09) |
| (*) | NLP_URJC | 65.77 | 88.89 (02) | 32.00 (07) | 47.06 (08) | 74.24 (08) | 98.00 (02) | 84.48 (04) |

Table 3: Spanish subtask leaderboard. The ranking is calculated using the Macro F1-score. We also include the Precision (P), Recall (R), and F1-score (F1) for the HOPE (HS) and Non-HOPE (NHS) labels. (*) This team sent their results out of time.

| | | | HS | | | NHS | | |
|---|---|---|---|---|---|---|---|---|
| # | **Team** | **Score** | P | R | F1 | P | R | F1 |
| 01 | I2C-Huelva | 50.12 | 1.63 (01) | 19.05 (04) | 3.01 (01) | 99.63 (04) | 94.96 (05) | 97.24 (05) |
| 02 | juanmanuel.calvo | 49.89 | 0.00 (06) | 0.00 (05) | 0.00 (06) | 99.56 (06) | 10.00 (01) | 99.78 (01) |
| 03 | Zavira | 49.75 | 0.00 (06) | 0.00 (05) | 0.00 (06) | 99.56 (07) | 99.44 (02) | 99.50 (02) |
| 04 | LIDOMA | 49.74 | 0.00 (06) | 0.00 (05) | 0.00 (06) | 99.56 (08) | 99.41 (03) | 99.49 (03) |
| 05 | NLP_SSN_CS | 49.37 | 0.00 (06) | 0.00 (05) | 0.00 (06) | 99.55 (09) | 97.95 (04) | 98.75 (04) |
| 06 | honghanhh | 48.62 | 1.28 (03) | 28.57 (03) | 2.46 (03) | 99.65 (02) | 90.36 (06) | 94.78 (06) |
| 07 | UMUTeam | 48.22 | 1.16 (04) | 28.57 (03) | 2.23 (04) | 99.65 (03) | 89.34 (07) | 94.21 (07) |
| 08 | mgraffg | 46.51 | 1.50 (02) | 61.90 (01) | 2.92 (02) | 99.80 (01) | 82.11 (08) | 90.09 (08) |
| 09 | Zootopi | 44.29 | 0.65 (05) | 33.33 (02) | 1.28 (05) | 99.62 (05) | 77.70 (09) | 87.30 (09) |
| (*) | NLP_URJC | 50.26 | 1.72 (01) | 19.05 (04) | 3.15 (01) | 99.63 (04) | 95.21 (05) | 97.37 (05) |
| (**) | Habesha | 48.94 | 1.54 (02) | 33.33 (02) | 2.94 (02) | 99.68 (04) | 90.64 (06) | 94.94 (06) |

Table 4: English subtask leaderboard. The ranking is calculated using the Macro F1-score. We also include the Precision (P), Recall (R), and F1-score (F1) for the HOPE (HS) and Non-HOPE (NHS) labels. (*) This team sent their results out of time. (**) This team forgot to publish the best result in the leaderboard.

task were far better than those of the English subtask, even though Spanish, as the Zootopi team pointed out, it is not a resource-rich language.

## *7 Conclusions and future work*

This paper presents the description of the first shared task on multilingual hope speech detection, organized within the IberLEF workshop, in the framework of the SEPLN 2023 conference. Specifically, two subtasks were proposed, the detection of HOPE speech in Spanish and English. Twelve different teams participated in the Spanish subtask and eleven in the English subtask. The best result from the Spanish subtask (Zootopi) achieved an average macro F1-score of 0.916 while in the English subtask it was 0.501 (I2C-Huelva). As future work, we pretend to improve the datasets by increasing their sizes, and reducing the unbalance among the labels in order to further promote the detection of hope speech.

## *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, Project FedDAP (PID2020-116118GA-I00) and Project Trust-ReDaS (PID2020-119478GB-I00) supported by MICINN/AEI/10.13039/501100011033, and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. It is also part of the research projects AIInFunds (PDC2021-121112-I00) and LTSWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

## *References*

Ahani, Z., G. Sidorov, O. Kolesnikova, and A. Gelbukh. 2023. Zavira at HOPE2023IberLEF: Hope Speech Detection from Text using TF-IDF Features andMachine Learning Algorithms. 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.*

Balaji, V., A. Kannan, A. Balaji, and B. Bharathi. 2023. NLP_SSN_CSE at HOPE2023IberLEF: Multilingual Hope Speech Detection using Machine Learning Algorithms. 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.*

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.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, (2020):1–10.

Cañete, J., S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo. 2022. Albeto and distilbeto: Lightweight spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298.

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), dec. Association for Computational Linguistics.

Chakravarthi, B. R., V. Muralidaran, R. Priyadharshini, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, M. A. García-Cumbreras, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumar Kumaresan, R. Ponnusamy, D. García-Baena, and J. A. García-Díaz. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. *Association for Computational Linguistics*, pages 378–388, may.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

*Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Domínguez Olmedo, J. L., J. Mata Vázquez, and V. Pachón Álvarez. 2023. I2C-Huelva at HOPE2023IberLEF: Simple Use of Transformers for Automatic 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*.

García-Baena, D., M. García-Cumbreras, S. M. Zafra, J. García-Díaz, and R. Valencia-García. 2023. Hope speech detection in spanish. *Language Resources and Evaluation*, pages 1–28, 03.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

Gemeda Yigezu, M., G. Yohannis Bade, O. Kolensikova, G. Sidorov, and A. Gelbukh. 2023. Multilingual Hope Speech Detection using Machine Learning. 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*.

He, P., X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Huertas-Tato, J., A. Martin, and D. Camacho. 2022. Bertuit: Understanding spanish language in twitter through a native transformer. *arXiv preprint arXiv:2204.03465*.

Jiménez-Zafra, S. M., F. Rangel, and M. Montes-y Gómez. 2023. 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*.

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*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Milne, D. N., G. Pink, B. Hachey, and R. A. Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 118–127.

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*.

no, A. G. F., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

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., G. Alcaraz-Mármol, and F. García-Sánchez. 2023. UMUTeam at HOPE2023IberLEF: Evaluation of Transformer Model with Data Augmentation for Multilingual 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*.

Rodríguez-García, M. A., A. Riaño Martínez, and S. Montalvo Herranz. 2023. URJC-Team at HOPE2023IberLEF: Multilingual Hope Speech Detection Using Transformers Architecture. 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.*

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Shahiki-Tash, M., J. Armenta-Segura, O. Kolesnikova, G. Sidorov, and A. Gelbukh. 2023. LIDOMA at HOPE2023IberLEF: Hope Speech Detection Using Lexical Features and Convolutional Neural Networks. 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.*