I2C-Huelva at HOPE2023@IberLEF: Simple Use of Transformers for Automatic Hope Speech Detection

Juan Luis Domínguez Olmedo, Jacinto Mata Vázquez and Victoria Pachón Álvarez

I2C Research Group, University of Huelva, Spain

Abstract

We present in this paper our participation in the share task Multilingual Hope Speech detection (HOPE) at IberLEF-2023. It consists of two binary subtasks with notable differences in terms of language and balance of labels. For the first task we have used BERTuit, a transformer model proposed for Spanish language; and for the second one we employed DistilBERT, a light transformer model trained by distilling BERT base. After several submissions for these tasks, we achieved an average macro F1 value of 0.744 in the first task, at the second position of the leaderboard; in the second task we finally ranked at first position, with an average macro F1 value of 0.501.

Keywords

Hope speech detection, Transformers, BERTuit, DistilBERT

1. Introduction

Hope Speech (HS) is the type of speech that can relax a hostile environment [1] and that helps, gives suggestions, and inspires for good to several people when they are in times of illness, stress, loneliness or depression [2].

By detecting it automatically, positive comments can be more widely disseminated, and can have a very significant effect when it comes to combating sexual or racial discrimination or when we seek to foster less bellicose environments [1].

Offensive messages on social media are posted towards people because of their race, color, ethnicity, gender, sexual orientation, nationality, or religion. As [2] stated, the importance of the social media lives of vulnerable groups, such as people belonging to the Lesbian, Gay, Bisexual, and Transgender (LGBT) community, racial minorities or people with disabilities, has been studied and it has been found that the social media activities of a vulnerable individual play an essential role in shaping the individual's personality and how he or she views society [3, 4, 5].

The shared task "HOPE. Multilingual Hope Speech detection" is part of IberLEF-2023, an evaluation campaign for Natural Language Processing (NLP) systems where several challenges are run with large international participation from research groups in academia and industry [6]. This 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. It consists of, given a text, written in Spanish or English, identifying whether it contains hope speech or not [7][8].

After our participation in this task, we present a summary of the work and the results obtained. The next section shows a description of the subtasks and datasets provided by the organizers. The experimental methodology and evaluation results are laid out in Sections 3 and 4. And some conclusions are presented in the last section.

IberLEF 2023, September 2023, Jaén, Spain

EMAIL: juan.dominguez@dti.uhu.es (J.L. Domínguez); mata@uhu.es (J. Mata); vpachon@uhu.es (V. Pachón)
ORCID: 0000-0001-5083-2313 (J.L. Domínguez); 0000-0001-5329-9622 (J. Mata); 0000-0003-0697-4044 (V. Pachón)

2. Description of the Subtasks and Datasets

Next, we will briefly describe the subtasks and datasets provided for the task "HOPE. Multilingual Hope Speech detection", part of IberLEF-2023.

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.

The data provided for this subtask consists of a set of LGBT-related tweets annotated as HS (Hope Speech) or NHS (Non-Hope Speech) [9].

A tweet is considered as HS if the text: i) explicitly supports the social integration of minorities; ii) is a positive inspiration for the LGTBI community; iii) explicitly encourages LGTBI people who might find themselves in a situation; or iv) unconditionally promotes tolerance. On the contrary, a tweet is marked as NHS if the text: i) expresses negative sentiment towards the LGTBI community; ii) explicitly seeks violence; or iii) uses gender-based insults.

The number of samples and distribution for each category in the training dataset provided by the organization is shown in Table 1. As it can be seen, the categories were almost equally distributed.

Table 1
Number of samples and distribution for each category in the training dataset (subtask 1)

Category	# samples	% of the total
NHS (Non-Hope Speech)	821	50.9 %
HS (Hope Speech)	791	49.1 %

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 English corpus provided for this subtask is an extension of the English part of the HopeEDI dataset [10]. It consists of comments posted on YouTube videos on a wide range of socially relevant topics such as Equality, Diversity, and Inclusion, including LGBTIQ issues, COVID-19, women in STEM, Black Lives Matter, etc.

The number of samples and distribution for each category in the training dataset provided by the organization is shown in Table 2. As it can be seen, there exist a clear unbalance in the dataset.

Table 2
Number of samples and distribution for each category in the training dataset (subtask 2)

Category	# samples	% of the total
NHS (Non-Hope Speech)	23221	91.2 %
HS (Hope Speech)	2229	8.8 %

2.3. Evaluation measures

To evaluate the results at both subtasks, precision, recall, and F1-score were measured per category and averaged using the macro-average method. Models were ranked using the macro-F1 score.

3. Methodology

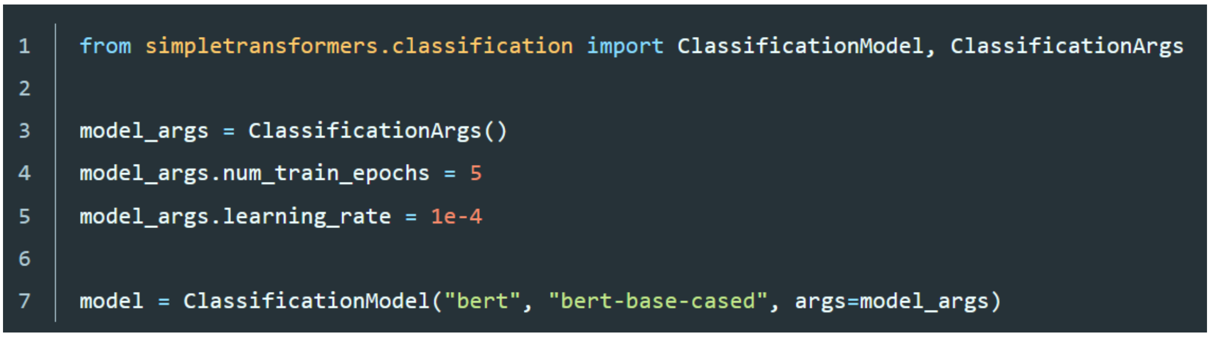
A Transformer is a deep learning model that adopts the self-attention mechanism, differentially weighting the importance of each part of the input data. It is frequently used in the fields of NLP and Computer Vision [11].

Simple Transformers is an NLP library designed to simplify the usage of transformer models without having to compromise on utility. It is built on the work of Hugging Face and their Transformers library [12, 13].

At the highest level, Simple Transformers is branched into common NLP tasks such as text classification, question answering, and language modeling. Each of these tasks have their own task-specific Simple Transformers model. It has built-in support for:

- Text Classification
- Token Classification
- Question Answering
- Language Modeling
- Language Generation
- Multi-Modal Classification
- Conversational AI
- Text Representation Generation

While all the task-specific models maintain a consistent usage pattern (initialize, train, evaluate, predict), this separation allows the freedom to adapt the models to their specific use case. Figure 1 shows the initial code for an example of a classification model.



```
from simpletransformers.classification import ClassificationModel, ClassificationArgs

model_args = ClassificationArgs()
model_args.num_train_epochs = 5
model_args.learning_rate = 1e-4

model = ClassificationModel("bert", "bert-base-cased", args=model_args)
```

Figure 1: Example of use of Simple Transformers for a classification model

Some of its benefits are the following:

- Input data formats are optimized for the task.
- Outputs are clean and ready-to-use for the task with minimal to no post-processing required.
- Unique configuration options for each task, while sharing a large, common base of configuration options across all tasks.

3.1. Subtask 1

For the “Hope Speech detection in Spanish” we have employed BERTuit, a transformer proposed for Spanish language, and pre-trained using RoBERTa optimization [14]. The transformer had been trained from scratch with text created by native speakers from Twitter, by using more than 230 million Tweets from the Archive Twitter Stream Grab [15], from 2021 to 2018.

The training dataset provided by the organizers consisted of 1612 records, of which 821 (51%) were of the NHS (Non-Hope Speech) category.

A basic pre-processing was applied to the text, consisting of:

- change text to lowercase
- remove http/https links
- remove the hash sign (#)
- strip whitespace (including newlines)

Apart from that, all the digits were eliminated from the text in some versions of the BERTuit model.

We have mainly used the default hyperparameters, some of which are shown in the Table 3.

Table 3
Some of the configuration options (hyperparameter default values)

Hyperparameter	Value
optimizer	AdamW
adam_epsilon	1e-8
learning_rate	4e-5
train_batch_size	8

3.2. Subtask 2

For the "Hope Speech detection in English" we have employed DistilBERT. It is a small, fast, cheap, and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark [16, 17].

The training dataset provided by the organizers consisted of 25450 records. After deleting those records with the same value in both 'text' and 'category', 24285 records remained, of which 22145 (91%) were of the NHS (Non-Hope Speech) category. It is clearly an unbalanced training dataset.

A basic pre-processing was applied to the text, consisting of:

- change text to lowercase
- remove http/https links
- remove the hash sign (#)
- remove the @ sign
- remove the emojis
- strip whitespace (including newlines)

Apart from that, some extra preprocessing was applied in some versions of the DistilBERT model. The additional preprocessing consisted in removing Unicode characters, and also removing single letters and numbers surrounded by space.

We have mainly used the default hyperparameters, some of which are shown in the Table 4.

Table 4
Some of the configuration options (hyperparameter default values)

Hyperparameter	Value
optimizer	AdamW
adam_epsilon	1e-8
learning_rate	4e-5
train_batch_size	8

4. Results

4.1. Subtask 1

Several versions of the BERTuit model were trained and tested using the test dataset provided by the organizers. The configuration options of the best ones are shown in Table 5.

Table 5
Some of the best submissions and their configuration options

Model	train_batch_size	# epochs	Digits removed
Subm-1	8	5	Not
Subm-2	16	5	Yes
Late submission	8	10	Yes

The corresponding evaluation metrics for these models are shown in Table 6. As it can be seen, the late submission improved the F1 measure for both categories, so it can be deduced that the digits removal worked well in this case.

Table 6
Evaluation metrics for Subtask 1

	Avg. Macro F1	Precision HS	Recall HS	F1 HS	Precision NHS	Recall NHS	F1 NHS
Subm-1	0.7437	**0.9091**	0.4667	0.6167	0.7855	**0.9767**	0.8707
Subm-2	0.7266	0.8919	0.4400	0.5893	0.7766	0.9733	0.8639
Late sub.	**0.7715**	0.8977	**0.5267**	**0.6639**	**0.8039**	0.9700	**0.8792**

4.2. Subtask 2

Several versions of the DistilBERT model were trained and tested using the test dataset provided by the organizers. The configuration options of the best ones are shown in Table 7.

Table 7
Some of the best submissions and their configuration options

Model	# epochs	Extra preprocessing
Subm-1	1	Not
Subm-2	2	Not
Late submission	1	Yes

The corresponding evaluation metrics for these models are shown in Table 8. As it can be seen, the late submission improved the macro F1, due to the increase in both the precision and recall of the HS category; so it can be deduced that the extra preprocessing worked well in this case.

Table 8
Evaluation metrics for Subtask 2

	Avg. Macro F1	Precision HS	Recall HS	F1 HS	Precision NHS	Recall NHS	F1 NHS
Subm-1	0.5012	0.0163	0.1905	0.0301	0.9963	**0.9496**	**0.9724**
Subm-2	0.4951	0.0163	**0.2857**	0.0309	0.9966	0.9245	0.9592
Late sub.	**0.5040**	**0.0212**	**0.2857**	**0.0395**	**0.9967**	0.9421	0.9686

5. Conclusions and Further Work

In this paper we present the methods and results for the shared task Multilingual Hope Speech detection (HOPE) at IberLEF-2023. We employed BERTuit and DistilBERT transformer models for the two binary subtasks, respectively. We have used the Simple Transformers library for the python language.

After some preprocessing and light tuning of hyperparameters, we achieved acceptable results in both subtasks after several submissions for testing. We ranked at the second position of the leaderboard in the Spanish language subtask, and at first position for the English language.

As future work that could improve the classification results, oversampling/undersampling techniques could be applied for Subtask 2, which presents a clear unbalance in the distribution of categories. Also, it might be worth further tuning of the hyperparameters.

Acknowledgements

This paper is part of the I+D+i Project titled "Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe".

We want to thank Juan M. Calvo Flores for his suggestion on the use of the Simple Transformers library.

References

[1] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, in: Proceedings of the 24th European Conference on Artificial Intelligence, ECAI 2020, IOS Press, 2020, pp. 226–236, doi:10.3233/FAIA200305.

[2] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, 2020, pp. 41–53.

[3] P. Burnap, G. Colombo, R. Amery, A. Hodorog, J. Scourfield, Multi-class machine classification of suicide-related communication on twitter, Online social networks and media 2 (2017) 32–44. doi:10.1016/j.osnem.2017.08.001.

[4] V. Kitzie, 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 (2018). doi: 10.5210/fm.v23i7.9264.

[5] D.N. Milne, G. Pink, B. Hachey, R.A. Calvo, in: Proceedings of the third workshop on computational linguistics and clinical psychology, 2016, pp. 118–127. doi:10.18653/v1/W16-0312.

[6] Jiménez-Zafra, S. M., Rangel, F., Montes-y-Gómez, M. (2023). Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, 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.

[7] IberLEF 2023 Task - HOPE. Multilingual Hope Speech detection, 2023. URL: https://codalab.lisn.upsaclay.fr/competitions/10215.

[8] Jiménez-Zafra, S. M., García-Cumbreras, M. Á., García-Baena, D., García-Díaz, J. A., Chakravarthi, B. R., Valencia-García, R., Ureña-López, L. A. (2023). Overview of HOPE at IberLEF 2023: Multilingual Hope Speech Detection. Procesamiento del Lenguaje Natural, vol 71, septiembre 2023.

[9] García-Baena, D., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., García-Díaz, J. A., & Valencia-García, R. (2023). Hope speech detection in Spanish: The LGBT case. Language Resources and Evaluation, 1-28.

[10] Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., Cn, S., McCrae, J. P., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., Valencia-García, R., Kumar Kumaresan, P., Ponnusamy, R., García-Baena, D. & García-Díaz, J. (2022). Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 378-388). https://aclanthology.org/2022.ltedi-1.58.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, pp. 5998-6008. URL: http://arxiv.org/abs/1706.03762.

[12] Simple Transformers. URL: https://simpletransformers.ai.

[13] Hugging Face. URL: https://huggingface.co.

[14] J. Huertas-Tato, A. Martin, D. Camacho, BERTuit: Understanding Spanish language in Twitter through a native transformer, 2022. URL: https://arxiv.org/abs/2204.03465.

[15] Twitter Stream Grab. URL: https://archive.org/details/twitterstream.

[16] DistilBERT. URL: https://huggingface.co/docs/transformers/model_doc/distilbert.

[17] V. Sanh, L. Debut, J. Chaumond, T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019. URL: https://arxiv.org/abs/1910.01108.