Silver Surfer team at Touché task 4: Testing Data Augmentation and Label Propagation for Multilingual Stance Detection*

Notebook for the Touché Lab on Argument and Causal Retrieval at CLEF 2023

Jorge Avila, Alvaro Rodrigo and Roberto Centeno

NLP & IR group at UNED, Madrid, Spain

Abstract

Touché task 4 evaluates systems performing stance detection in a multilingual setting where a reduced annotated dataset is available. We have tested different approaches focused on increasing training data by (1) including new samples from back-translating original training data and (2) adding automatically annotated data using label propagation. According to the results, back-translation was quite successful in improving results, with our best baselines using it. On the other hand, with label propagation, we obtained worse results than without using it. The current results, close to a 0.35 f1 score, show that there is still room for improvement in this task.

Keywords

Data augmentation, Label propagation, Multilingual Stance detection

1. Introduction

The Touché Lab at CLEF proposes a series of shared tasks focused on computational argumentation and causality [1]. In this paper, we focused on our participation in task 4: Intra-Multilingual Multi-Target Stance Classification. The objective of this task is to classify comments on socially relevant topics that have been written on the Conference on the Future of Europe (CoFE)[1] platform. CoFE is an online platform where any user can write a proposal in any of the 24 official languages of the EU. Other users can comment on and/or endorse a proposal or another comment. The task is to classify whether these comments are in favor, against, or neutral toward the proposal. The proposals, titles, and comments can be written in any of the 24 languages of the European Union.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

javila149@alumno.uned.es (J. Avila); alvarory@lsi.uned.es (A. Rodrigo); rcenteno@lsi.uned.es (R. Centeno)

http://nlp.uned.es/~alvarory/ (A. Rodrigo)

0000-0002-6331-4117 (A. Rodrigo); 0000-0001-9095-4665 (R. Centeno)

[1]https://futureu.europa.eu/?locale=en

Early tasks on stance detection, like those at SemEval-2016 Task 6 [2], only provided texts in a single language, usually English. Then, more recently, new initiatives proposed stance detection in other languages and include additional data. For example, SardiStance@EVALITA2020 proposes to detect stance about the Sardines movement[2] in Italian tweets, including contextual information related to users [3]. Afterward, VaxxStance@IberLEF 2021 launched a shared task in Spanish and Basque for detecting stance towards vaccines [4], including also information related to the social network. These two tasks showed the importance of considering users' information when detecting the stance of a given text. However, all these tasks focused on monolingual stance detection about single topics. Thus, Touché, where comments are written in different languages, represents a real challenge given the multilingual and multi-topic nature of the data [5, 6].

The data provided in the development period is mainly divided into three subsets:

- CF_S: contains 7000 comments annotated using only two classes (favor or against)
- CF_U: contains 12000 unlabeled comments
- CF_E-Dev: contains 1400 multilingual comments annotated with three classes.

One of the main difficulties of this task is the small size of the 3-class subset, which follows the schema required for annotating the test set. This is why we have explored different alternatives to exploit the information from the other two subsets, which were unlabeled or labeled using only two classes. Thus, our main objective was to expand the data used for training our models. For this purpose, we mainly rely on data augmentation and label propagation.

The rest of the paper is structured as follows: Section 2 describes the method for adding training data using data augmentation, while label propagation is described in Section 3. Then, we described the runs submitted in Section 4, analyzing their results in Section 5. Finally, some conclusions and future work are given in Section 6.

2. Data Augmentation

Data augmentation is a technique used to increase the quantity and diversity of training data. It involves applying transformations or modifications to existing data to generate new instances that are different but still contain the same information or labels [7].

In this work, we have applied a data augmentation method based on back-translation [8]. This method is based on generating variants of the original text in different languages and then translating them back to the original language to enrich the original dataset.

We wanted to explore this approach because, although it is better to expand the training data with completely different messages, this has associated a high annotation cost. By using data augmentation, we can automatically produce new data that is slightly different from the original one.

[2]https://en.wikipedia.org/wiki/Sardines_movement

3. Label Propagation

Label propagation is a semi-supervised machine-learning technique that can be used to propagate known labels onto unlabeled data [9]. The main objective is to utilize the information available in labeled data to assign labels to unlabeled data [10].

Label propagation is particularly useful in situations where there is a limited set of labeled data but a large amount of unlabeled data available. For example, in this task, the development dataset contains only 1400 comments, while the organizers provided unlabeled data with 12000 comments. Thus, label propagation is suitable for increasing our training data using the unlabeled data provided by the organizers.

We have firstly represented all the comments from the labeled and unlabeled subsets (respectively CF_E-Dev and CF_U subsets) into an embedding space. For this purpose, we have used the *paraphrase-multilingual-mpnet-base-v2*[3] model from the HuggingFace API[4] was used to embed the comments. This model is trained over 50 languages, making it suitable for the multilingual data we have.

Then, we have applied the LabelSpreading[5] algorithm from scikit-learn. This algorithm builds a similarity matrix including regularization, which is more robust to noise.

4. Submitted Runs

We have submitted six different runs to test different approaches using the TIRA platform [11]. Five of these runs are based on training Bert-base models [12], which is a common approach for similar classification tasks. We have selected these runs based on previous experiments using cross-validation on the CF_E dataset provided in the development period. The runs submitted are:

- Run 1: an XGBoost model [13] feed with metadata and the output of 6 models trained on different alterations of the provided datasets. More in detail, this run uses metadata such as the number of up/downvotes, endorsements, etc, and the output probabilities from the following systems:
  - A RoBERTa base model[6] trained on CF_E translated into English (run 5).
  - A XLM-RoBERTa large model[7] trained on CF_E.
  - An XLM-RoBERTa large model[8] trained on CF_S.
  - A RoBERTa base model[9] trained on CF_S translated into English.
  - Run 6.

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[4] https://huggingface.co/inference-api
[5] https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html#sklearn.semi_supervised.LabelSpreading
[6] https://huggingface.co/roberta-base
[7] https://huggingface.co/xlm-roberta-large
[8] https://huggingface.co/xlm-roberta-large
[9] https://huggingface.co/roberta-base

  - Similar to run 6, but using a BERT multilingual base (uncased)[10] model (so the comments are not translated into English).

- Run 2: A RoBERTa base model[11] trained on CF_E adding the CF_U dataset (translated into English) after applying the label propagation method described in Section 3. With this run, we wanted to study the effect of label propagation for stance detection.
- Run 3: A XLM-RoBERTa large model[12] trained on CF_E adding the CF_U dataset after applying the label propagation method described in Section 3. In this run, we wanted to study the influence of language on label propagation in the previous run, where we translated comments into English.
- Run 4. An XLM-RoBERTa large model[13] trained on the CF_E dataset augmented using the back-translation method described in Section 2. This run aimed to study the effect of data augmentation on stance detection.
- Run 5: A RoBERTa base model[14] using the CF_E subset translated into English. We consider this run as our baseline for comparing results with those using label propagation or data augmentation.
- Run 6. A BERT base model (uncased)[15] fine-tuned in 2 steps. First step: fine-tuning using the CF_S subset translated into English. Second step: fine-tuning using the CF_E dataset translated into English. With this run, we wanted to test the effect of transferring learning from a task with a bigger dataset annotated with two classes, to stance detection using three classes

The complete list of hyperparameters are given in Appendix A.

5. Analysis of Results

We show in Table 1 the results of our runs and the baseline proposed by the organizers. Results are sorted by the official measure, macro f1 score.

We can see in Table 1 how all the runs, except run 3, outperform the proposed baseline. All the results are under 0.4, showing that there is still room for improvement in this task. However, we obtained in our best experiments at the development period a score of 0.6691 with run 1. So, we think that our models were unable to correctly generalize the training data.

The best results are obtained by run 6, showing the importance of using additional data for fine-tuning the model, even if this data uses a different number of labels. Besides, the good results of run 4, the second in the ranking, also shows the importance of including additional training data, obtained for this run using back-translation. We also have similar results with run 1, which uses an ensemble of classifiers trained on different datasets. Hence, it seems quite promising to use approaches based on generating additional training data. All these runs outperformed run 5, which is considered our baseline.

[10] https://huggingface.co/bert-base-multilingual-uncased
[11] https://huggingface.co/roberta-base
[12] https://huggingface.co/xlm-roberta-large
[13] https://huggingface.co/xlm-roberta-large
[14] https://huggingface.co/roberta-base
[15] https://huggingface.co/bert-base-uncased

Table 1
Results of the submitted runs sorted by macro f1

Run	macro f1-score	macro precision	macro recall
Run 6	0.35	0.597	0.354
Run 4	0.329	0.582	0.328
Run 1	0.323	0.629	0.299
Run 5	0.27	0.552	0.255
Run 2	0.239	0.511	0.238
Touche23-baseline	0.237	0.851	0.333
Run 3	0.216	0.5	0.21

However, the results using label propagation for the unlabeled collection were not so successful as we can see with the results of run 3, the only run worse than the baseline, and run 2. Both runs performed worse than run 5, which only uses the CF_E subset. Therefore, we need to further research how to properly take advantage of unlabeled data for this task.

6. Conclusions and Future Work

Stance detection is a widely used task to understand the opinion or attitude expressed in texts. It is applied in a variety of contexts, such as analyzing product or service reviews, monitoring social media, and identifying online comments on public interest topics. The Touché Lab proposes a task for stance detection in a multilingual environment, with a diverse set of topics.

Given the nature of the task, with a reduced set of labeled data, we have studied different approaches focused on adding training data to feed our models. More in detail, we have mainly tested two approaches: (1) data augmentation using back-translation of the development set and (2), label propagation of the unlabeled data provided by the organizers.

Our best systems were those using the augmented data generated using back-translation, outperforming a similar model only using the available labeled data. However, our runs that used the unlabeled data did not perform so well. Hence, additional training data seems to be important for this task, but we need to further research how to properly generate this kind of data.

Acknowledgments

This work has been partially funded by the Spanish Research Agency (Agencia Estatal de Investigación), DeepInfo project PID2021-127777OB-C22 (MCIU/AEI/FEDER,UE) and the HOLISTIC ANALYSIS OF ORGANISED MISINFORMATION ACTIVITY IN SOCIAL NETWORKS project (PCI2022-135026-2).

References

[1] A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. Reimer, B. Stein, M. Potthast, M. Hagen, Overview of Touché 2023: Argument and Causal Retrieval, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, p. to appear.

[2] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, Semeval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 31–41.

[3] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, R. Paolo, et al., Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets, in: EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, CEUR, 2020, pp. 1–10.

[4] R. Agerri, R. Centeno, M. S. Espinosa, J. F. de Landa, Á. Rodrigo, Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection, Proces. del Leng. Natural 67 (2021) 173–181. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6387/3807.

[5] V. Barriere, G. G. Jacquet, L. Hemamou, Cofe: A new dataset of intra-multilingual multi-target stance classification from an online european participatory democracy platform, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp. 418–422.

[6] V. Barriere, A. Balahur, Multilingual multi-target stance recognition in online public consultations, Mathematics 11 (2023) 2161.

[7] D. A. Van Dyk, X.-L. Meng, The art of data augmentation, Journal of Computational and Graphical Statistics 10 (2001) 1–50.

[8] A. Sugiyama, N. Yoshinaga, Data augmentation using back-translation for context-aware neural machine translation, in: Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 35–44. URL: https://aclanthology.org/D19-6504. doi:10.18653/v1/D19-6504.

[9] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Label propagation for deep semi-supervised learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5070–5079.

[10] F. Wang, C. Zhang, Label propagation through linear neighborhoods, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 985–992.

[11] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://doi.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[13] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. URL: https://doi.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.

A. Hyperparameters

A.1. Run 1

- subsample: 0.5
- min_child_weight: 5
- max_depth: 7
- learning rate: 0.01
- colsample_bytree: 0.5

A.2. Run 2

- Batch size: 14
- Leaning rate: 1 x 10^{-5}
- Weight decay: 0.001
- Epochs: 6

A.3. Run 3

- Batch size: 2 (with accumulation of gradients on 4 batches)
- Leaning rate:1 x 10^{-5}
- Weight decay: 0.001
- Epochs: 5

A.4. Run 4

- Batch size: 2
- Leaning rate: 1 x 10^{-5}
- Weight decay: 0.001
- Epochs: 8

A.5. Run 5

- Batch size: 8
- Leaning rate: 6 x 10^{-6}
- Weight decay: 0.001
- Epochs: 6

A.6. Run 6

- Batch size: 8
- Leaning rate: 1 x 10^{-5}
- Weight decay: 0.001
- Epochs: 3