Unsupervised Ranking and Aggregation of Label Descriptions for Zero-Shot Classifiers

Angelo Basile
angelo.basile@symanto.com
Symanto Research
Universitat Politècnica de València
Spain

Marc Franco-Salvador
marc.franco@symanto.com
Symanto Research
Spain

Paolo Rosso
prosso@dsic.upv.es
PRHLT Research Center
Universitat Politècnica de València
Spain

Abstract
Zero-shot text classifiers based on label descriptions embed an input text and a set of labels into the same space: measures such as cosine similarity can then be used to select the most similar label description to the input text as the predicted label. In a true zero-shot setup, designing good label descriptions is challenging because no development set is available. Inspired by the literature on Learning with Disagreements, we look at how probabilistic models of repeated rating analysis can be used for selecting the best label descriptions in an unsupervised fashion. We evaluate our method on a set of diverse datasets and tasks (sentiment, topic and stance). Furthermore, we show that multiple, noisy label descriptions can be aggregated to boost the performance.

1 Introduction
Recently, large Language Models (LMs) such as BERT (Devlin et al., 2019) have pushed the boundaries of NLP systems and have enabled a transition from the supervised learning paradigm, where an input text is processed together with a ground-truth label, to a pre-train, prompt and predict paradigm (Liu et al., 2021), where a pre-trained LM is fed the input data to be processed and a description of the task to be performed. This paradigm shift has lead to zero-shot models that require no ground-truth labels. With a good task description, zero-shot models have been shown to be effective at many challenging NLP tasks (Brown et al., 2020). However, LMs are highly sensitive to how a task description is framed (Jiang et al., 2020) and, without a large development set, finding a good task description is hard. In this work we address this problem, focusing on zero-shot models based on Siamese BERT-Networks (SBERT) (Reimers and Gurevych, 2019). These networks embed both the input text and a description of the target labels in the same semantic space using a pre-trained LM; in this space, similarity measures are then applied to map the most similar label description to the most probable labels. As with prompting, a good label description is key to obtaining good performance. For instance, in the context of binary sentiment analysis, the words awesome, perfect, great and bad, terrible, awful are all potentially good descriptions for the labels positive and negative, respectively.

How do we filter out the sub-optimal descriptions without having access to labelled data? We study the Learning with Disagreements literature (Uma et al., 2021), particularly, the item-response class of models and show that methods developed for analysing crowd-sourced annotations can be transferred to the problem of description selection for zero-shot models. It has been shown that these models achieve high performance in two tasks related to disagreement analysis. First, they usually outperform majority voting at retrieving the gold truth. Second, they can identify which annotators are more reliable and which are spammers. In this work, we look at how we can use these models in the domain of model-based zero-shot classification. Figure 1 shows an illustration of the proposed method.

2 Zero-Shot Classification with SBERT
Zero-Shot classification models can tackle a task using no training data. Starting from a pre-trained LM, different architectures enable zero-shot classification using different strategies, such as prompting (Brown et al., 2020; Liu et al., 2021) or Natural Language Inference (NLI) for label entailment (Yin et al., 2020). In this work, we focus on zero-shot classifiers based on Siamese Networks. They have recently been shown to perform on par with other methods while being highly efficient at inference time (Müller et al., 2022; Chu et al., 2021). As
Figure 1: Overview of the proposed method for a sentiment analysis task with two possible output labels, negative (NEG) and positive (POS). An unlabeled corpus with $T$ documents is embedded through $a$ SBERT together with a set of $n$ label descriptions $L$. A softmax classifier on top of the SBERT cosine similarity scores provides $n$ predicted labels $y$, one for each item in the label description set. These predictions are then passed as input to $b$ the probabilistic aggregator, which outputs $\hat{y}$, a best guess of a single final predicted label for each document $T$ and $\theta$, a reliability score for the whole corpus for each label description $y$.

Label Descriptions
Label descriptions are the key component that make it possible to turn a semantic similarity model into a zero-shot classifier. We experiment with four sets of label descriptions. First, we define as a baseline a identity hypothesis (IH) label, which we set to be equal to the class name it describes: for example, in the context of sentiment analysis, we define the words positive and negative as label descriptions for the class POSITIVE and NEGATIVE, respectively. Second, as a first source of variation, we experiment with a set of patterns for turning a identity hypothesis label into a proper sentence: for example, for the IMDB movie review dataset, we experiment with the pattern The movie is $\{\text{positive}, \text{negative}\}$. Third, for each dataset we manually write multiple variations over the identity hypothesis (e.g., $\{\text{positive}, \text{negative}\} \rightarrow \{\text{great}, \text{terrible}\}$). Finally, we experiment with automatically generated variations of the identity hypothesis: under the assumption that the representation of a word in different languages can be a useful source of additional information for a multilingual encoder, we use a machine translation system to automatically translate the identity hypothesis labels into three different languages, which can be directly fed to a multilingual pretrained SBERT model (see Section 4 for more details). Table 1 shows the label descriptions used for the IMDB movie review dataset.

3 Bayesian Label Description Analysis
From Crowdsourcing to Zero-Shot Classifiers. Bayesian inference provides a natural framework for dealing with multiple sources of uncertain information. This framework is successfully used in the context of analysing crowdsourced annotations as a robust alternative to: i) inter-annotator agreement metrics for identifying biases and potential reliability issues in the annotation process, and ii) majority voting for retrieving the gold truth label. In this work, we argue that this framework can be directly applied to the problem of prompt and label descrip-
Table 1: Overview of the label descriptions used for the IMDB movie review dataset. The identity hypothesis (IH) labels (positive, negative) describe the class positive and negative respectively. The rows MANUAL and PATTERN have been manually compiled, while row AUTO shows the label descriptions generated automatically by translating the null hypothesis.

|          | POSITIVE | NEGATIVE |
|----------|----------|----------|
| **IH**   | positive | negative |
|          | great    | terrible |
| **MANUAL** | really great | really terrible |
|          | a masterpiece | awful |
| **AUTO**   | optimo   | terrible |
|          | grande   | terivel  |
|          | genial   | negativo |

Figure 2: The model plate for MACE. Given $I$ instances and $N$ annotators, the observed label $y_{i,n}$ is dependent on the gold label $G_i$ and the behaviour $B_{i,n}$ of the zero-shot model with label description $i$ on the instance $n$. The variable $B$ was originally introduced in MACE for modelling the spamming behaviour of crowd-workers. Key of this work is that, in the context of zero-shot classifiers, $B$ correlates strongly with the true f1-score of different label descriptions. As a consequence, it can be used to rank the label descriptions and eventually discard the sub-optimal ones. We refer to the original MACE paper (Hovy et al., 2013) for additional details on the model.

### Table 2: Spearman $\rho$ rank correlation with true f1-score for both the $\theta$ parameter of the MACE model and the baseline Cohen’s $\kappa$ score.

| Dataset        | $\rho(\theta, f1)$ | $\rho(\kappa, f1)$ |
|----------------|---------------------|---------------------|
| Ag News        | 0.58                | 0.38                |
| Cola           | 0.12                | 0.08                |
| Imdb           | 0.94                | 0.87                |
| StanceCat      | 0.66                | 0.39                |
| SubJ           | 0.09                | **0.20**            |
| Yelp           | **0.83**            | 0.76                |
| Yelp Full      | 0.16                | **0.43**            |

In NLP, a popular Bayesian annotation model is the Multi-Announcer Competence Estimation (MACE) model (Hovy et al., 2013). MACE is a generative, unpooled model. It is generative because it can generate a dataset starting from a set of priors, and the assumptions that produce a specific outcome can be represented as a graph, as shown in Figure 2. Thanks to its unpooled structure, it models each annotator (or label description) independently and provides a trustworthiness score $\theta_j$ for each annotator $j$: we use this parameter for ranking the label descriptions assuming that high $\theta$ values lead to higher f1-scores. Given $n$ label descriptions and $i$ instances, the zero-shot model outputs $n$ predicted labels $y$: MACE models each label $y_{i,n}$ for instance $i$ from a label description $n$ as being dependent on the true, unobserved gold label $G_i$ and the behaviour $B_{i,n}$ of the zero-shot model with label description $i$ on the instance $n$. The variable $B$ was originally introduced in MACE for modelling the spamming behaviour of crowd-workers. Key of this work is that, in the context of zero-shot classifiers, $B$ correlates strongly with the true f1-score of different label descriptions. As a consequence, it can be used to rank the label descriptions and eventually discard the sub-optimal ones. We refer to the original MACE paper (Hovy et al., 2013) for additional details on the model.

### 4 Experiments and Results

#### 4.1 Setup

For implementing the zero-shot classification module, we use the Python package `sentence-transformers` (Reimers and Gurevych, 2019). For our experiments on the English corpora, we use the pre-trained model `paraphrase-MiniLM-L3-v2` (Reimers and Gurevych, 2021), which has been trained on a variety of different datasets. We evaluate our proposed method on a battery of popular text classification datasets: IMDB (Maas et al., 2011), Yelp Review (Zhang et al., 2015), AG’s News Corpus (Gulli, 2005), Cola (Warstadt et al., 2018). In addition, we include StanceCat (Taulé et al., 2017), a stance detection dataset for which non-aggregated annotations are available. For the identity hypothesis (IH) label descriptions, we re-use the prompts that we could find in the literature on prompting (Yin et al., 2019; Wang...
et al., 2021; Müller et al., 2022) and manually crafted the rest. \( ^1\) We manually wrote the variations on the identity hypothesis based on our intuitions. For the automatic generation of new label descriptions, we translated the English identity hypothesis into French, Italian and Spanish using a pre-trained MarianMT model (Junczys-Dowmunt et al., 2018) through the transformers library (Wolf et al., 2020). Specifically, we used the opus-mt-en-roa model (Tiedemann and Thottingal, 2020).

We conduct our experiments using the Stan-based implementation (Paun et al., 2018) of MACE, trained with VB.

### 4.2 Results

**Ranking.** Table 2 shows the rank correlation scores with the true f1-score for different datasets. As a baseline, we use the average of Cohen’s \( \kappa \) (Artstein and Poesio, 2008) computed between each pair of label descriptions. For most of the datasets, the MACE’s \( \theta \) parameter, which models the trustworthiness of an annotator, outperforms the baseline. Medium to strong correlation between MACE’s \( \theta \) parameters and the true f1-score suggests that a model-based analysis using zero-shot classifiers can be used to effectively select the best performing label descriptions and discard the sub-optimal ones, i.e., by ranking the different label descriptions according to the \( \theta \) values, the low-scoring labels can be safely left out.

**Aggregation.** Table 3 shows the results of the label aggregation experiments. In all of the cases, aggregating multiple label descriptions outperforms the identity hypothesis (IH) baseline. In addition, MACE usually outperforms majority voting, excluding the cases where the label space contains more than two labels (i.e., Ag News and Yelp Full). On average, the automatically generated label descriptions outperform both the manually written label descriptions and the pattern variations: this suggests that human involvement is not necessarily needed for finding better performing label descriptions.

### 5 Related Work

The idea of using the meaning of a category for building dataless classification systems has first been explored already in pre-neural times (Chang et al., 2008). Within the pretrain, prompt and predict paradigm (Liu et al., 2021), automatic prompt generation and ensembling has been investigated in Jiang et al. (2020). Schick and Schütze (2021) train a classifier on top of the soft-labels provided by an ensemble of zero-shot models for successfully cancelling the effect of poorly-performing prompts. The idea of modelling ambiguity and disagreement in annotation as signal more than noise, has recently gained traction in the NLP community (Uma et al., 2021; Plank et al., 2014; Fornaciari et al., 2021). The closest source to our paper is probably (Simpson et al., 2013), who use a Bayesian model to combine multiple weak classifiers in a better performing system. (Paun et al., 2018; Passonneau and Carpenter, 2014) highlight the benefits of Bayesian models for NLP specifically.

### 6 Conclusion

We set out to address two research questions: first, we looked at the problem of unsupervised ranking of different label descriptions by their estimated trustworthiness on different text classification datasets; second, we investigated whether the

| method | aggregation | Ag News | Cola | Imdb | StanceCat | SubJ | Yelp | Yelp Full |
|--------|-------------|---------|------|------|-----------|------|------|----------|
| IH     | -           | 10.9    | 38.7 | 66.2 | 32.4      | 52.1 | 66.8 | 31.3     |
| PATTERN | mace        | 8.2     | 41.9 | 68.1 | 30.8      | 54.5 | 73.3 | 7.4      |
|        | majority    | 8.1     | 40.6 | 67.2 | 31.4      | 51.7 | 72.3 | 34.7     |
| MANUAL | mace        | 9.7     | 38.7 | 66.3 | 33.1      | 55.2 | 69.8 | 33.5     |
|        | majority    | 9.9     | 38.7 | 66.2 | 28.0      | 50.0 | 68.3 | 31.0     |
| AUTO   | mace        | 10.3    | 47.3 | 68.7 | 31.5      | 43.3 | 74.9 | 7.4      |
|        | majority    | 24.8    | 47.1 | 66.5 | 31.0      | 45.7 | 72.5 | 34.0     |

Table 3: Macro-averaged F1 scores for the experiments with label aggregations.

\( ^1\)The complete of the label descriptions used can be found at https://github.com/anbasile/zsla/
output of zero-shot models built with different label descriptions can be aggregated in order to obtain overall higher classification performance. We found that Bayesian models of annotations such as MACE can provide a good solution for both problems. Furthermore, we have found that automatically translated label descriptions outperform manually written ones. We focused on Siamese zero-shot models because their inference runtime is not affected by the number of label descriptions. When put all together, these findings suggest that zero-shot model performance can be improved by automatically generating more label descriptions and aggregating their output with a probabilistic model.

Acknowledgements

We gracefully thank the support of the Pro²Haters - Proactive Profiling of Hate Speech Spreaders (CDTi IDI-20210776), XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), DETEMP - Early Detection of Depression Detection in Social Media (IVACE IMINOD/2021/72) and DeepPattern (PROMETEO/2019/121) R&D grants. Grant PLEC2021-007681 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4):555–596.

Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. Advances in NIPS, pages 1877–1901.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, et al. 2008. Importance of semantic representation: Dataless classification. In Aaai, pages 830–835.

Zewei Chu, Karl Stratos, and Kevin Gimpel. 2021. Unsupervised label refinement improves dataless text classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4165–4178, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2591–2597, Online. Association for Computational Linguistics.

Antonio Gulli. 2005. AG’s corpus of news articles. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? Transactions of the Association for Computational Linguistics, 8:423–438.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Bich, 2018. Marian: Fast neural machine translation in C++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, et al. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ArXiv.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Müller, Guillermo Pérez-Torró, and Marc Franco-Salvador. 2022. Few-Shot Learning with Siamese Networks and Label Tuning. In ACL.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. Transactions of the Association for Computational Linguistics, 2:311–326.
