Abstract
This paper presents our solution for SemEval-2022 Task 10: Structured Sentiment Analysis. The solution consisted of two modules: the first for sequence tagging and the second for relation classification. In both modules we used transformer-based language models. In addition to utilizing language models specific to each of the five competition languages, we also adopted multilingual models. This approach allowed us to apply the solution to both monolingual and cross-lingual sub-tasks, where we obtained average Sentiment Graph F1 of 54.5% and 53.1%, respectively. The source code of the prepared solution is available at https://github.com/rafalposwiata/structured-sentiment-analysis.

1 Introduction
Structured Sentiment Analysis (SSA) can be formulated as an information extraction task in which one attempts to find all of the opinion tuples \( O = O_1, ..., O_n \) in a text. Each opinion \( O_i \) is a tuple \((h, t, e, p)\) where \( h \) is a holder who expresses a polarity \( p \) towards a target \( t \) through a sentiment expression \( e \), implicitly defining pairwise relationships between elements of the same tuple (Barnes et al., 2021). An example of such tuples as a structure sentiment graph was shown in Figure 1. This problem is relatively new and there has been little work published on the subject to date. To stimulate interest in this issue among the NLP community the SemEval-2022 Task 10: Structured Sentiment Analysis (Barnes et al., 2022) competition was organized. The contest consisted of two sub-tasks: monolingual and cross-lingual. In the monolingual sub-task, the systems were trained and then tested on the datasets in the same languages. In the cross-lingual sub-task, systems had to be prepared for Catalan, Basque and Spanish datasets, while data in these languages could not be used for training. This setup is often known as zero-shot cross-lingual transfer (Hu et al., 2020).

In this paper we present our system for this competition. We mainly focused on the solution for the monolingual track, however, it has also been successfully applied to the cross-lingual. The rest of the paper is organized as follows. Section 2 briefly describes related work. Section 3 shows an overview of used datasets. Section 4 elaborates on our solution. Experiments showing the effectiveness of the created system performed on development and test sets are presented in Section 5. The next section briefly describes the mistakes and limitations of our system. Finally, Section 7 concludes this paper.

2 Related Work
Structured Sentiment Analysis can be broken down into five sub-tasks: a) expression (opinion) extraction, b) target (aspect) extraction, c) holder extraction, d) defining the relationship between these elements, and e) assigning polarity (Barnes et al., 2021).2

A few years ago, the main focus was on Aspect-Based Sentiment Analysis (ABSA), which only concerned on targets extraction (task b) and classifying the polarity towards them (task e) (Pontiki et al., 2014, 2015, 2016). Sequence tagging solutions have proven to be effective in this issue (Li et al., 2019a). An extension of this problem was End2End Aspect-Based Sentiment Analysis (E2E-ABSA), which adds the issue of expression extraction (task a). He et al. (2019) propose an interactive multi-task learning network (IMN) which is able to jointly learn multiple related tasks simultaneously, to resolve this problem. Chen and Qian (2020) also use multi-task learning, but with relation propagation mechanisms and create Relation-Aware Collaborative Learning (RACL) framework. Tagging-based solutions also work well in this case.

1 Picture based on figure from Barnes et al. 2021.
2 Phrases in parentheses indicate alternative names used interchangeably in the sentiment analysis literature.
Some others give the new UMUC 5 stars - don’t believe them.

Figure 1: SSA example as a structure sentiment graph.  

| Dataset | # sentences | # tags |
|---------|-------------|-------|
|         | all w/o opinion | w/ one opinion | w/ two or more opinions | w/ mixed tags | w/ nested tags | w/ opposite polarity exp. |
| MPQA   | train 5873 | 4619 | 917 | 337 | 92 | 108 | 0 | 1425 | 1481 | 698 | 337 | 671 |
|         | dev 2063 | 1647 | 304 | 112 | 49 | 38 | 0 | 406 | 494 | 215 | 124 | 231 |
|         | test 2113 | 1724 | 289 | 100 | 31 | 36 | 0 | 434 | 462 | 229 | 124 | 165 |
| DSUnis | train 2253 | 1572 | 583 | 98 | 3 | 0 | 1 | 63 | 806 | 364 | 102 | 340 |
|         | dev 232 | 150 | 69 | 13 | 0 | 0 | 0 | 9 | 98 | 54 | 15 | 29 |
|         | test 318 | 214 | 84 | 20 | 0 | 0 | 0 | 12 | 130 | 62 | 12 | 56 |
| OpeNERen | train 1744 | 344 | 638 | 762 | 0 | 0 | 0 | 266 | 2679 | 783 | 0 | 2101 |
|         | dev 249 | 51 | 83 | 115 | 0 | 0 | 0 | 49 | 371 | 116 | 0 | 284 |
|         | test 499 | 92 | 178 | 229 | 0 | 0 | 0 | 98 | 793 | 269 | 0 | 596 |
| OpeNERes | train 1438 | 186 | 500 | 752 | 0 | 0 | 0 | 176 | 2748 | 570 | 0 | 2472 |
|         | dev 206 | 32 | 77 | 97 | 0 | 0 | 0 | 23 | 363 | 70 | 0 | 317 |
|         | test 410 | 48 | 159 | 203 | 0 | 0 | 0 | 56 | 849 | 189 | 0 | 768 |
| MultiBen | train 1174 | 172 | 508 | 494 | 0 | 0 | 0 | 169 | 1705 | 716 | 0 | 1273 |
|         | dev 167 | 27 | 79 | 61 | 0 | 0 | 0 | 15 | 211 | 107 | 0 | 151 |
|         | test 335 | 54 | 143 | 138 | 0 | 0 | 0 | 53 | 434 | 204 | 0 | 319 |
| MultiBca | train 1063 | 164 | 478 | 421 | 0 | 0 | 0 | 205 | 1277 | 278 | 0 | 1401 |
|         | dev 152 | 32 | 68 | 52 | 0 | 0 | 0 | 33 | 152 | 36 | 0 | 167 |
|         | test 305 | 65 | 126 | 114 | 0 | 0 | 0 | 58 | 331 | 65 | 0 | 372 |
| NoRecFine | train 8634 | 4079 | 2406 | 2149 | 802 | 472 | 173 | 898 | 6778 | 2753 | 0 | 5695 |
|         | dev 1531 | 710 | 441 | 380 | 119 | 87 | 32 | 120 | 1152 | 444 | 0 | 988 |
|         | test 1272 | 598 | 353 | 321 | 123 | 79 | 14 | 110 | 993 | 359 | 0 | 876 |

Table 1: Statistics of the datasets. Mixed tags means a situation where a given term in different opinions plays a different role, e.g. once it is a target and once it is a holder. Nested tags are when a term in one opinion is part of a term in another opinion. Opposite polarity expressions refers to the case where a sentence contains an expression that has a different sentiment depending on the opinion.

The recently proposed, Aspect Sentiment Triplet Extraction (ASTE) fill this gap (Peng et al., 2020). The task is to extracting all aspects terms with their corresponding opinion terms and sentiment polarity (tasks a, b, d and e). Peng et al. (2020) propose two stage model. In the first stage, it extracts opinions and aspects along with sentiment using sequence tagging based on the unified BIO scheme. The second stage pairs up the predicted terms from the first stage to output triplets. ASTE is most similar to SSA, missing only the holder extraction.

For SSA, the subject of the competition, there are few solutions. Barnes et al. (2021) cast the structured sentiment problem as dependency graph parsing. Peng et al. (2021) extend this work and propose a sparse and fuzzy attention scorer with pooling layers which improves parser performance.

3 Datasets

Seven structured sentiment datasets in five languages were selected for the competition. The MPQA dataset (Wiebe et al., 2005) contains news documents from the world press in English. DSUnis (Toprak et al., 2010) are English reviews of online universities and e-commerce. OpeNERen and OpeNERes (Agerri et al., 2013) consist of hotel reviews in English and Spanish, respectively. MultiBen and MultiBca (Barnes et al., 2018) are also hotel reviews, but in Basque and Catalan. The last dataset is NoRecFine (Øvrelid et al., 2020), a multi-domain dataset of professional reviews in Norwegian. The statistics of each dataset are sum-
Some others give the new UMUC 5 stars - don’t believe them.

Transformer-based
Sequence Tagging
(Extraction Module)

R-BERT
Relation Classifier
(Relation Classification Module)

Create Pairs
Expression
Target / Holder
Some others
5 stars
the new UMUC
Some others
don’t believe
the new UMUC

Create Output
Expression Polarity Target Holder
R-BERT

Figure 2: Architecture of the proposed solution.

4 System Overview

The architecture of our solution is shown in Figure 2. This solution was inspired by the works of Li et al. (2019a, b); Hu et al. (2019), and especially the work of Peng et al. (2020). It consists of two main components: Extraction Module and Relation Classification Module. The first module is based on sequence tagging and is used to extract targets, holders and expressions with polarity. This is accomplished by using a suitable tagset which is a modification of the BIO scheme, consisting of the following tags: {B-holder, B-targ, B-exp-Neg, B-exp-Neu, B-exp-Pos, I-holder, I-targ, I-exp-Neg, I-exp-Neu, I-exp-Pos, O}. Transformer-based Language Model with a linear classification layer was used as an implementation. Having already extracted entities, the role of the second module is to classify whether there is a relationship between them. Specifically, it is about verifying that there is a holder and/or target associated with a particular expression. We utilized the R-BERT (Wu and He, 2019) model to accomplish this task. Based on a sentence with two appropriately marked entities (expression and holder/target), it determines whether or not they are related. For all the details, we would refer you to Wu and He 2019 paper.

5 Experiments

5.1 Experimental Setup

To conduct the experiments, we first utilized the Simple Transformers library (Rajapakse, 2019) for the implementation of the Extraction Module. For the Relation Classification Module we modify publicly available source code of R-BERT. The hyperparameters used in learning each of these modules are presented in Table 2. All models were run five times on a single GPU Tesla V100.

3https://github.com/monologg/R-BERT
Table 2: Parameter used for Extraction and Relation Classification modules during training.

| Parameter          | Extraction | Relation Classification |
|--------------------|------------|-------------------------|
| Optimizer          | AdamW      | AdamW                   |
| Learning rate      | 5e-5       | 2e-5                    |
| Batch size         | 32         | 16                      |
| Dropout            | 0.1        | 0.1                     |
| Epochs             | 10         | 12                      |
| Validation after no. steps | 200   | 200                     |

5.2 Pretrained Language Models

We chose two types of language models based on transformer architecture for experiments: monolingual (at least one for each of the five competition languages) and multilingual. The use of multilingual models allowed us to obtain a more general solution and was necessary for the cross-lingual sub-task. Table 3 gives a brief summary of the models used. All models were downloaded from the Hugging Face hub.

| Language | Model       | Size | Source                   |
|----------|-------------|------|--------------------------|
| English  | BERT        | base | Devlin et al. 2019       |
|          | RoBERTa     | large| Liu et al. 2019          |
|          | XLNet       | large| Yang et al. 2019         |
| Spanish  | BERTIN      | base | de la Rosa et al. 2021   |
|          | RoBERTa-BNE | large| Gutiérrez-Fandiño et al. 2021 |
| Catalan  | Catalan-BERTa| base | Armengol-Estapé et al. 2021 |
| Basque   | BERTeus     | base | Agerri et al. 2020       |
| Norwegian| NB-BERT     | base | Kutzov et al. 2021       |
|          | mBERT       | base | Devlin et al. 2019       |
|          | XLM-R       | large| Conneau et al. 2020      |

Table 3: Transformer-based language models used in experiments.

5.3 Metrics

Following the works on Named Entity Recognition problem (Akbik et al., 2018; Yamada et al., 2020; Zhou and Chen, 2021), we used micro-average F1 score as our main measure for the Extraction Module. In addition for this module we added a detailed measure for each tag type i.e. F1 score for holders, targets and expressions with sentiment classes, separately. For the Relation Classification Module, we used Accuracy and macro-average F1 measures. Evaluation of the overall system was based on the official competition metric i.e. Sentiment Graph F1.

5.4 Development Results

Table 4 shows the results on the development sets for each module. For the Extraction Module, the XLM-R model was the best on five of the seven datasets. In only two cases (MPQA and DSUnis) language-specific models were found to be superior: XLNet and RoBERTa, respectively. For the Relation Classification Module, we only used models based on the BERT architecture, following the original R-BERT work (Wu and He, 2019). The mBERT usually proved to be the best (5/7 cases), except for two cases (MultiB and NoReCFine) where BERTeus and NB-BERT were the best. The best models for each module were used to test the overall system. A summary of this experiment can be found in Table 5. The average Sentiment Graph F1 was 55.0%.

5.5 Test Results

The best models verified on the development sets were used on the test sets which are the official competition sets. For the monolingual sub-task, we used exactly the same configuration of models as in Table 5. For the cross-lingual sub-task, we used models trained on the OpeNERen set, namely XLM-R for extraction and mBERT for relation classification. There were two reasons for this choice. First is the use of multilingual models in both modules. Second, from the fact that the results on the development sets were high compared to the results for other models trained on English language sets. The results are summarized in Table 6. We achieved average SF1 scores of 54.5% and 53.1% for the monolingual and cross-lingual sub-tasks, respectively. This allowed us to rank 11th and 9th out of the 32 teams in these sub-tasks.

6 Errors Analysis

As a result of the used architecture, most errors are due to incorrect tagging. In particular, this is relevant to expressions where a correct sentiment is additionally required. The results were significantly worse for expressions limited in a given set, e.g., neutrals in the MPQA or DSUnis sets. Furthermore, by using a single extraction model, the solution is not able to correctly handle more complicated cases such as mixed or nested tags or opposite polarity expressions. This is most noticeable in the NoReCFine dataset.
### Table 4: Results for the Extraction and Relation Classification modules on development sets. Underlined and bolded numbers indicate the best result for the metric and dataset.

| Dataset | Extraction | Relation Classification |
|---------|------------|--------------------------|
| MPQA    | BERT       | 50.4                     |
|         | RoBERTa    | 58.8                     |
|         | XLNet      | 57.9                     |
|         | mBERT      | 49.3                     |
|         | XLM-R      | 56.8                     |
| DSUnis  | BERT       | 22.2                     |
|         | RoBERTa    | 50.0                     |
|         | XLNet      | 66.7                     |
|         | mBERT      | 18.2                     |
|         | XLM-R      | 28.6                     |
| OpeNERen| BERT       | 71.6                     |
|         | RoBERTa    | 71.4                     |
|         | XLNet      | 66.7                     |
|         | mBERT      | 66.7                     |
|         | XLM-R      | 75.0                     |
| OpeNERes| BERTIN     | 77.4                     |
|         | RoBERTa-BNE| 71.4                     |
|         | mBERT      | 66.7                     |
|         | XLM-R      | 75.0                     |
| MultiBea| Catalan-BERTa | 69.2                    |
|         | RoBERTa-BNE| 52.6                     |
|         | mBERT      | 61.5                     |
|         | XLM-R      | 67.0                     |
| MultiBea| BERTena    | 61.2                     |
|         | RoBERTa-BNE| 59.8                     |
|         | mBERT      | 64.1                     |
|         | XLM-R      | 69.4                     |
| NoRecCeva| NorBERT     | 62.0                     |
|         | NB-BERT    | 64.6                     |
|         | mBERT      | 54.7                     |
|         | XLM-R      | 63.4                     |

### Table 5: Overall system results on development sets.

| Dataset | Monolingual | Cross-lingual |
|---------|-------------|---------------|
| MPQA    | 32.6        | -             |
| DSUnis  | 39.5        | -             |
| OpeNERen| 67.0        | -             |
| OpeNERes| 66.3        | 56.4          |
| MultiBea| 65.0        | 58.6          |
| NoRecCeva| 45.9        | -             |

### Table 6: Overall system results on test sets (official results of the competition).

| Dataset | Monolingual | Cross-lingual |
|---------|-------------|---------------|
| MPQA    | 32.6        | -             |
| DSUnis  | 39.5        | -             |
| OpeNERen| 67.0        | -             |
| OpeNERes| 66.3        | 56.4          |
| MultiBea| 65.0        | 58.6          |
| NoRecCeva| 45.9        | -             |

### 7 Conclusion

In this paper, we presented a solution to the SemEval-2022 Task 10: Structured Sentiment Analysis. A simple architecture based on sequence tagging and relation classification achieved good results. The use of multilingual language models enabled the solution to be used for monolingual and cross-lingual sub-tasks. At the same time it can be easily extended e.g. by using an additional CRF layer (Souza et al., 2019) in the Extraction...
Module or by using other multilingual language models e.g. InfoXLM (Chi et al., 2021).

References
Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. In Sociedad Española para el Procesamiento del Lenguaje Natural, volume 51, pages 215–218.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Aitor Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representations some love: the case for basque. In Proceedings of the 12th International Conference on Language Resources and Evaluation.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodríguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4933–4946, Online. Association for Computational Linguistics.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Òvrelið, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Oberländer Laura Ana Maria Kutuzov, Andrey and, Enrica Troiano, Jan Buchmann, Rodrigo Agerri, Lilja Òvrelið, Erik Velldal, and Stephan Oepen. 2022. SemEval-2022 task 10: Structured sentiment analysis. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Seattle. Association for Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3685–3694, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin 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, Online. Association for Computational Linguistics.

Javier de la Rosa, Eduardo González, Pablo Villegas, Pablo González de Prado, Manu Romero, and María Grandury. 2021. BERTIN project.

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.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodríguez-Penagos, and Marta Villegas. 2021. Spanish language models.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. CoRR, abs/2003.11080.

Minghao Hu, Xuying Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 537–546, Florence, Italy. Association for Computational Linguistics.
