Multi-Label Classification of Patronizing and Condescending Language
ANLP Project Report

Laura Riedel
University of Potsdam / Advanced Natural Language Processing 2021/2022
laura.riedel@uni-potsdam.de

Abstract

Detecting and classifying the types of patronising and condescending language (PCL) is an important task of our time to reduce the use of and educate about unintentionally harmful language. In the context of the SemEval 2022 Task 4 (Pérez Almendros et al., 2021) we investigate the performance differences of a Logistic Regression and a bidirectional LSTM model when training and classifying on different levels of abstraction. While we cannot recommend our model types to be used for PCL classification in applied settings, we find that PCL classification generally seems to profit from a higher level of abstraction.

1 Introduction

Patronising and condescending language (PCL) is a type of language use that communicates a superior attitude towards vulnerable social groups or depicts them in a pitying way (Pérez Almendros et al., 2020a). This discriminatory effect is usually unconscious with the author having good intentions (Pérez Almendros et al., 2020a); nevertheless this kind of language use contributes to routinising and obscuring discrimination (Ng, 2007), thus impeding vulnerable communities in their attempt to overcome the difficulties they already face and reaching total inclusion (Nolan and Mikami, 2013). The detection of PCL is therefore imperative for reducing the use of potentially harming language in our society. Being able to classify which kind of PCL is being used would furthermore add to raising awareness for the problems of PCL use and educating about what kind of PCL one encounters or unknowingly produces.

In this project we approach PCL as a multi-label classification task. We choose to implement two model types – a logistic regression and a bidirectional LSTM – with the goal of comparing how the classification performances change with the models' diverging levels of advance. We furthermore investigate the topic from two levels of abstraction. By modelling the BiLSTM's architecture closely to an implementation that has been carried out previously on the same dataset by current professionals of the field we hope to be able to evaluate our models more realistically in the light of those contrasting results.

2 Related Work

Our proposal builds directly on Pérez Almendros et al. (2020a) which is one of the trailblazing works regarding patronising and condescending language (PCL) from an NLP perspective. In their paper, Pérez Almendros et al. introduce a newly created dataset annotated for different kinds of PCL, and they also illustrate successes and challenges of PCL classifiers by testing out different kinds of NLP models on their dataset.

For proper classification, Pérez Almendros et al. (2020a) developed a taxonomy of PCL types and their relations. They identified seven distinct categories of PCL, which in turn are grouped into three higher-level categories:

- The saviour:
  - Unbalanced power relations
  - Shallow solution
- The expert:
  - Presupposition
  - Authority voice
- The poet:
  - Metaphor
  - Compassion
  - The poorer, the merrier

Pérez Almendros et al. (2020a) came to the conclusion that identifying PCL is hard for standard NLP models and that models like BERT tended to perform better. However, they found that categories like *Presupposition* that require some form of world knowledge seem to present an obstacle to models more generally.

In order to raise awareness for this subtle type of discrimination and promote the development of NLP models for this key problem, the authors presented the same two PCL detection tasks they themselves tackled (Pérez Almendros et al., 2020a) as task in the SemEval 2022 challenge (Pérez Almendros et al., 2021). This SemEval task serves as our starting point, as will be elaborated in the next section.

3 Problem Statement

We are taking on the SemEval 2022 Task 4 Subtask 2 which deals with multi-label classification of PCL in the following way: “Given a paragraph, a system must identify which PCL categories express the condescension” (Pérez Almendros et al., 2021). The labels used for classification comply with the taxonomy proposed by (Pérez Almendros et al., 2020a) explained in Section 2. For example, the given dataset (further details see Section 4) contains the paragraph

> “Chantelle Owens, Mrs Planet 2016, hosted the day and the ladies had the chance to share their compassion for those in need.”

for which it provides the lower-level gold labels *Unbalanced power relations*, *Shallow solution*, and *Compassion*.

While the SemEval task only asks to classify paragraphs according to their lower-level labels, we are extending this task by including the taxonomy’s higher-level labels as alternative label set. We consider two model types with three settings each and compare them to a random baseline. Crucially, we assume that each given paragraph indeed contains some instance of PCL.

3.1 Task Formalisation

The task is set out as a multi-label classification task with a collection P of k paragraphs $\{p_0, p_1, ..., p_k\}$ and a set L of n labels $\{l_0, l_1, ..., l_n\}$. Each paragraph $p \in P$ is represented by z-dimensional embedding vectors. Depending on whether we intend on classifying lower-level (*ll*) or higher-level (*hl*) labels, the number of labels changes: $n = 7$ for L_{ll} and $n = 3$ for L_{hl}, respectively. The goal is for a model to classify each paragraph $p \in P$ with between 1 and n labels $l \in L$.

4 Data

We use the version of the Don’t Patronize Me! dataset (Pérez Almendros et al., 2020a) which the authors provided as training set for the SemEval 2022 competition. The dataset is divided into two parts; the first listing all paragraphs and stating whether or not they contain an instance of patronizing and condescending language (PCL), the second consisting only of those paragraphs actually containing PCL and being labelled for the types of PCL appearing according to the authors’ own PCL taxonomy. For the multi-label classification task of PCL detection we thus use the latter part of the dataset providing the PCL labels. In this light, the original purpose of the dataset perfectly mirrors the task at hand.

Don’t Patronize Me! consists of selected paragraphs of news stories published between 2010 and 2018 in twenty English speaking countries that were extracted from the News on Web corpus (Davies, 2013). Paragraphs were selected based on a list of ten keywords targeting certain vulnerable groups and concepts often used to describe the former (Pérez Almendros et al., 2020b): disabled, homeless, hopeless, immigrant, in-need, migrant, poor families, refugees, vulnerable and women.

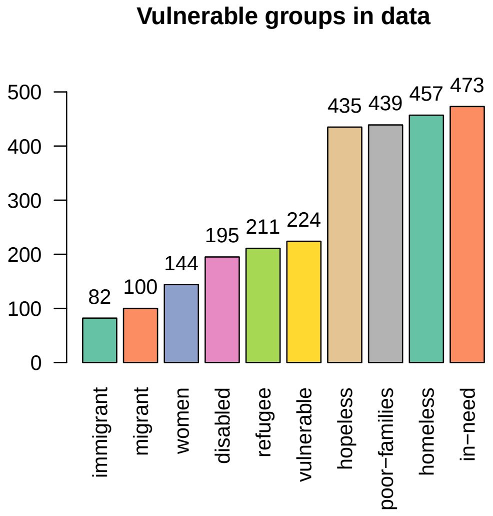

Figure 1: Distribution of vulnerable groups in paragraphs containing instances of PCL.

Out of the original collection of paragraphs only 993 contain instances of PCL which means we have relatively little data at our disposal for our models to train on. This is aggravated by the fact that the test set provided for the SemEval 2022 challange does not contain labels, meaning we implemented our own train/test split based on the provided SemEval train data. It is interesting to note that while the original paragraphs were selected in a balanced manner for country and keyword, the paragraphs actually containing PCL are less balanced concerning the vulnerable groups (See Figure 1) and terribly unbalanced with regards to the PCL categories (see Figure 2)[1].

The data were annotated by three female annotators with backgrounds in communication, media and data science who have Spanish as a first language but who are bilingual in English (Pérez Almendros et al., 2020b). Since the annotators present a very homogeneous group the annotations might be biased, especially towards the vulnerable group of women. For a description of how inter-annotator disagreements were resolved please see Pérez Almendros et al. (2020a).

It is important to point out that the results on Pérez Almendros et al.’s 2020 paper cannot be replicated as their current version of the dataset is an updated, more realistic version of the original dataset with different numbers (C. Pères Almendros, personal communication, March 16, 2022). The data statement provided on the SemEval 2022 github page would suggest that we are thus working with version 1.1. of the dataset (Pérez Almendros et al., 2020b) – however, the amount of paragraphs claimed by the data statement to be in the dataset do not match the amount of paragraphs listed in the provided dataset itself. While the data

[1] Plots were created using R (R Core Team, 2020).

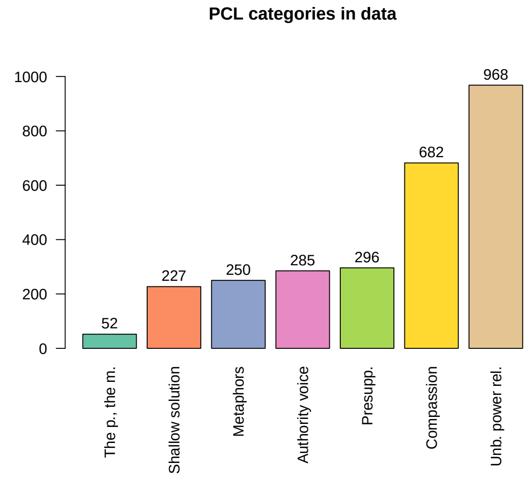

Figure 2: Distribution of types of PCL in paragraphs containing instances of PCL.

statement claims the Don't Patronize Me! dataset to consist of 7,638 paragraphs (Pérez Almendros et al., 2020b), `dontpatronizeme_pcl.tsv` lists 10,469 paragraphs in total, which would fit better as a training split to the 10,637 paragraphs the original paper lists (Pérez Almendros et al., 2020a). Therefore, it remains unclear which version of the Don't Patronize Me! dataset we are actually provided with.

5 Pre-Processing

In order to be able to use the Don't Patronize Me! dataset's paragraphs as inputs to our different models, we convert each $p \in P$ into model-specific embeddings: tf-idf sentence embeddings for the Logistic Regression and pre-trained word embeddings for the BiLSTM (details see Section 6). For the BiLSTM we prepare the input by applying our own `inp2array()` function to each $p \in P$ which converts the input paragraph (a string) into a list of word indices. In that process the function splits tokens still containing punctuation (e.g. 'poverty-stricken', 'life-changing', 'refugee/migrant', 'women.Greg') into two parts in order to increase token recognisability, and it ignores tokens that are not part of the pre-trained word embedding vocabulary. This effectively reduces the paragraph lengths to the amount of words we can infer meaning from with the given vocabulary. Furthermore, since it is useful for the input sequences to a BiLSTM to have the same length, we pad each $p \in P$ to a uniform length. In order to avoid unnecessary padding for the vast majority of sequences, we remove one outlier paragraph that is more than twice as long as the second longest paragraph from the dataframe generated from the dataset (i.e. none of the models interact with it). We also convert the ll labels $l \in L_{ll}$ given in the dataset into multi-hot encoded vectors and additionally infer (multi-hot encoded) hl labels $l \in L_{hl}$ from $l \in L_{ll}$. Finally, as mentioned in Section 4, we implement our own 80/20 train/test split based on the SemEval 2022 Don't Patronize Me! train data; references to train and test data allude to our own split from here on. The dataset is split according to a multi-label stratified sampling of the PCL categories using scikit-multilearn's *IterativeStratification* module (Szymański and Kajdanowicz, 2017).

6 Experiments

There are two main experimental settings: Classifying a given paragraph $p \in P$ according to lower-level (ll) labels $l \in L_{ll}$ and classifying according to higher-level (hl) labels $l \in L_{hl}$ of PCL instances. For this we consider two model types: Logistic Regression as a simpler probabilistic classifier and a bidirectional Long Short-Term Memory model (BiLSTM) as a more advanced approach for classification. We are curious to see what a model as simple as a Logistic Regression classifier can achieve in comparison to more complex models; particularly considering that Logistic Regression cannot make use of contextual information. A BiLSTM, on the other hand, is capable of 'remembering' important information from sequential data – both from the left and right context window due to its bidirectionality – which is a useful feature considering our rather pragmatic task.

Of the model types Logistic Regression and BiLSTM we consider two versions each, a 'simpler' version which we use for both experimental settings, and a k-fold cross-validation version which has the potential of alleviating the fact that we have so little data at our disposal. We will only implement k-fold cross-validation in the ll setting. We model our BiLSTM as closely as possible to Pérez Almendros et al. (2020a)'s BiLSTM so as to increase meaningful comparability across this model type to their results. In this section we first illuminate the embeddings employed in Section 6.1. After that we describe the specific model architectures used in the experiments in Section 6.2. Finally, which evaluation metrics we use to assess our models' performance and learning progress will be explained in Section 6.3.

6.1 Embeddings

For the Logistic Regression models we convert each $p \in P$ into a single tf-idf sentence embedding using scikit-learn's `TfidfVectorizer` with each embedding having the shape $(1, 6859)$.

For the BiLSTM we use the 300 dimensional word2vec Skip-gram word embeddings trained on the Google News dataset created by Mikolov et al. (2013). Combined with our padding (see Section 5) each $p \in P$ thus has the shape $(191, 300)$.

6.2 Models

Random Baseline. Our baseline model should not be entirely random but, in order to present a more meaningful comparison, reflect the distribution of labels found in the dataset. For this reason we use the underlying label distribution to generate weighted random multi-hot

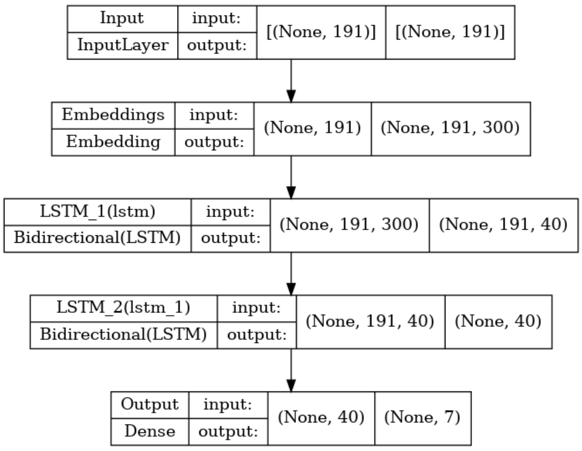

Figure 3: Illustration of the simple BiLSTM architecture for the classification of lower-level labels.

label encodings for $l \in L_{ll}$ and $l \in L_{hl}$ for each $p \in P$ in the test data.

Logistic Regression: simple. We implement a Logistic Regression model using scikit-learn's `LogisticRegression` with a `MultiOutputClassifier` wrapper (Pedregosa et al., 2011) and (paragraph-based) sentence embeddings as explained in Section 6.1. We set the maximum number of iterations to 500 but kept the default settings otherwise. With the multi-output wrapper this model is able to predict all $l \in L_{ll}$ and $l \in L_{hl}$ independently of each other.

Logistic Regression: k-fold cross-validation. This Logistic Regression has the same build as the simpler Logistic Regression mentioned above with the addition of a cross-validation wrapper around it. For that we use `KFold` from scikit-learn (Pedregosa et al., 2011) with 10 folds. The 10-fold cross-validation model is only employed for predicting $l \in L_{ll}$.

BiLSTM: simple. We implement a bidirectional Long Short-Term Memory model (BiLSTM) using the open-source library Keras (Chollet et al., 2015) with word embeddings as explained in Section 6.1. The model has one input layer, one embedding layer, two bidirectional LSTM layers, and a dense output layer. The BiLSTM layers have a hidden size of 20 and a dropout rate of 0.25% each. We train for 300 epochs using the Adam optimiser, with a batch size of 32, early stopping monitoring binary cross-entropy loss and a patience of 10 epochs. For the output layer we use a sigmoid activation function and an output size fitting the number of labels, $n = 7$ for $l \in L_{ll}$ and $n = 3$ for $l \in L_{hl}$.

BiLSTM: k-fold cross-validation. Similar to the cross-validation of Logistic Regression earlier, this bidirectional LSTM has the same build as the simpler BiLSTM mentioned above with the addition of a cross-validation wrapper around it. Again we use `KFold` from scikit-learn (Pedregosa et al., 2011) with 10 folds. The 10-fold cross-validation model is only employed for predicting $l \in L_{ll}$.

6.3 Evaluation

To make the training process of our models more tracable we collect the cross-entropy loss and accuracy values of each epoch during training. In the k-fold cross-validation setting we additionally collect the mean loss, accuracy, and F1-measure per fold and choose the fold with the best mean F1-measure as best-performing model for further evaluation.

For evaluating the models' performances on the test set we calculate each model's mean F1-score. We also calculate each model's accuracy, precision, recall, and F1-measure and display a confusion matrix for each of the n labels. This enables us to assess the strengths and shortcomings of the models more clearly and put the mean F1-score into perspective.

7 Results

In this section we present our model performances in two parts: First regarding lower-level (ll) classification in Section 7.1, then regarding higher-level (hl) classification in Section 7.2.

7.1 Lower-level Classification

In this section we begin by first looking at each simple model's individual results, then comparing them with the Random Baseline and drawing an interim conclusion. After that, we will consider each k-fold cross-validation model's individual results before comparing the performance of all ll-classifying models overall.

Logistic Regression: simple. As can be seen in Table 1, while the Logistic Regression has a high percentage of correct predictions for each $l \in L_{ll}$, it would be a fallacy to conclude that this model actually performs well. The confusion matrix (Table 14 in A) shows that four labels *Shallow solution, Presupposition, Metaphor,* and *The poorer, the merrier* are categorically predicted to not occur, and *Autority voice* is predicted to occur only once. The opposite trend is visible for *Unbalanced power relations*, this label is almost always predicted to be present. The only more balanced confusion matrix is the one for *Compassion*. Comparing these results to

	A	P	R	F1
Unbal. power rel.	80.23	80.68	99.30	89.02
Shallow solu.	77.97	0.00	0.00	0.00
Presupposition	74.58	0.00	0.00	0.00
Authority voice	74.58	100	2.17	4.26
Metaphor	77.97	0.00	0.00	0.00
Compassion	66.67	78.69	51.06	61.94
The p., the mer.	94.35	0.00	0.00	0.00

Table 1: Accuracy, precision, recall, and F1 score for each lower-level label predicted by multi-label Logistic Regression.

the distribution of ll-labels in the dataset (see Figure 2) shows that this prediction pattern seems to be directly related to the amount of label appearances in the corpus with *Unbalanced power relations* and *Compassion* being approximately thrice and twice as frequent as the other five labels, respectively.

BiLSTM: simple. The BiLSTM does not predict categorically for or against a single label, which is a nice improvement to the simple Logistic Regression (see Table 14). However, the label *The poorer, the merrier* is still conspicuous in that the model predicted an abundance of true negatives – leading to an excellent accuracy value (see Table 2) – but not a single true positive. The F1-scores are again best for those labels that appear most often in the dataset: *Unbalanced power relations* and *Compassion*. All others stay below 50% (see Table 2).

	A	P	R	F1
Unbal. power rel.	80.23	90.30	84.62	87.36
Shallow solu.	76.27	45.71	41.03	43.24
Presupposition	76.27	57.14	26.67	36.36
Authority voice	75.71	54.55	39.13	45.57
Metaphor	75.71	41.67	25.64	31.75
Compassion	68.36	77.94	56.38	65.43
The p., the mer.	93.22	0.00	0.00	0.00

Table 2: Accuracy, precision, recall, and F1 score for each lower-level label predicted by a BiLSTM.

Comparing the F1-scores of the simple Logistic Regression and BiLSTM to our Random Baseline shows that the BiLSTM performs best in four of the seven labels $l \in L_{ll}$ (see Table 3). The Logistic Regression model outperforms the BiLSTM by almost 2% for *Unbalanced power relations*, but this is likely simply due to lucky guessing on the Logistic Regression's part. Both the Logistic Regression and the BiLSTM model are dwarfed by our Random Baseline regarding the labels *Presupposition* and *The poorer, the merrier*. The latter can be somewhat excused by its incredibly low number of appearances (52 in the entire dataset, see Figure 2). It remains to be seen if the predictions for *Presupposition* will improve in our k-fold cross-validation setting.

	Random	LogReg	BiLSTM
Unbal. power rel.	75.29	**89.02**	87.36
Shallow solu.	24.00	0.00	**43.24**
Presupposition	**38.20**	0.00	36.36
Authority voice	19.75	4.26	**45.57**
Metaphor	29.33	0.00	**31.75**
Compassion	47.67	61.94	**65.43**
The p., the mer.	**12.5**	0.00	0.00

Table 3: Comparing F1 scores for each lower-level label from Random Baseline, Logistic Regression, and BiLSTM.

Logistic Regression: k-fold cross-validation. The metric values for the Logistic Regression with k-fold cross-validation are identical to the ones for the simple Logistic Regression in most cases. Table 4 shows only a minor difference for the label *Compassion* where the precision value is marginally higher for the cross-validation model and the other three values are marginally worse compared to the simple model. Accordingly, the confusion matrix also only changes slightly for *Compassion* (see Table 14).

	A	P	R	F1
Unbal. power rel.	80.23	80.68	99.30	89.02
Shallow solu.	77.97	0.00	0.00	0.00
Presupposition	74.58	0.00	0.00	0.00
Authority voice	74.58	100	2.17	4.26
Metaphor	77.97	0.00	0.00	0.00
Compassion	65.53	80.0	46.81	59.06
The p., the mer.	94.35	0.00	0.00	0.00

Table 4: Accuracy, precision, recall, and F1 score for each lower-level label predicted by multi-label Logistic Regression with 10-fold cross validation.

BiLSTM: k-fold cross-validation. Our k-fold cross-validation BiLSTM performs worse on the test data than the simple BiLSTM in all categories except for *Authority voice* (see Table 5). This comes as a surprise to us since we were expecting the cross-validation to alleviate the impact of relatively small numbers of positive instances and the unbalanced distribution of categories in the underlying dataset somewhat, thus improving the label predictions. For the three labels *Unbalanced power relations, Presupposition* and *Compassion* the F1 score generally decreases, for *Shallow solution* the precision value increases a little while recall and F1 measure decrease, for *Metaphor* the recall value stays the same while precision and F1 measure decrease. For *Authority voice* the precision and recall values seem to converge in the middle compared to their values with the simple BiLSTM. *The poorer, the merrier* stays entirely the same. Looking at the confusion matrices in Table 14 we can see that the way the simple and the cross-validation BiLSTMs predict does not differ that greatly:

	A	P	R	F1
Unbal. power rel.	77.97	90.0	81.82	85.71
Shallow solu.	77.40	48.39	38.46	42.86
Presupposition	72.32	41.67	22.22	28.99
Authority voice	72.32	46.67	45.65	46.15
Metaphor	75.14	40.0	25.64	31.25
Compassion	59.32	63.41	55.32	59.09
The p., the mer.	93.22	0.00	0.00	0.00

Table 5: Accuracy, precision, recall, and F1 score for each lower-level label predicted by a BiLSTM with 10-fold cross validation.

except for *Compassion*, the number of correct and false label predictions mostly changes by very few instances. This is mirrored in the performance differences mostly being relatively small. For example, for *Unbalanced power relations*, the cross-validation BiLSTM predicts four paragraphs falsely negative where in the simple BiLSTM there were four more true positive label predictions which on paper constitutes a difference of 1.65 percentage points.

Overall, very little changes from the trends we already saw with the simpler models. As can be seen in Table 6, the Random Baseline still has the best F1-scores for the labels *Presupposition* and *The poorer, the merrier*. Both Logistic Regression models reach an equally high F1-score for *Unbalanced power relations* with their rather categorical prediction of this label. The simple BiLSTM keeps achieving the best results in all other categories except for *Authority voice* where it is outperformed by the k-fold cross-validation BiLSTM. It is striking, however, that only for two of those $l \in L_{ll}$ the best F1 measures scored higher than 50%, namely *Unbalanced power relations* and *Compassion*.

	Ran.	LR	LR V	Bi-L.	Bi-L. V
Unba.	75.29	**89.02**	**89.02**	87.36	85.71
Shal.	24.00	0.00	0.00	**43.24**	42.86
Pres.	**38.20**	0.00	0.00	36.36	28.99
Auth.	19.75	4.26	4.26	45.57	**46.15**
Meta.	29.33	0.00	0.00	**31.75**	31.25
Comp.	47.67	61.94	59.06	**65.43**	59.09
Tptm.	**12.5**	0.00	0.00	0.00	0.00

Table 6: Comparing F1 scores for each lower-level label from Random Baseline, simple Logistic Regression, cross-validated Logistic Regression, simple BiLSTM, and cross-validated BiLSTM.

7.2 Higher-level Classification

In this section we start with looking at each simple model's individual results predicting on hl-labels before comparing them with the Random Baseline.

Logistic Regression: simple. Just like in the ll classification section, Logistic Regression gets fairly high accuracy values for each of the given labels (see Table 7). However, the scores for precision, recall and F1 improved drastically for $l \in L_{hl}$ compared to $l \in L_{ll}$: *The saviour* excelling in all scores is not terribly surprising since it contains the ever-present *Unbalanced power relations* category. Looking at the confusion matrix in Table 15 shows that the model predicted not a single true negative and only one false negative, indicating again towards a simple overgeneralisation for this label based on its omnipresence. While both *Presupposition* and *Authority voice* were predicted terribly in both ll Logistic Regressions before (see Tables 1 and 4), their shared hl label *The expert* improved dramatically (see Table 7) from an F1-score close to 0% to 39.18%. The hl label *The poet* is comprised of the three ll labels *Metaphor, Compassion* and *The poorer, the merrier*. While the previous Logistic Regression models performed relatively well on *Compassion*, they had F1-scores of 0% for the other two ll labels in this category. Insofar, the F1-score for this hl label improved quite a bit too in comparison to its ll components. Looking at the confusion matrix also reveals that the Logistic Regression model seems to have picked up on patterns more confidently for *The poet*, and there are far more true positives and true negatives than false ones.

	A	P	R	F1
The saviour	84.18	84.67	99.33	91.14
The expert	66.67	79.17	26.03	39.18
The poet	70.62	74.11	78.30	76.15

Table 7: Accuracy, precision, recall, and F1 score for each higher-level label predicted by a simple multi-label Logistic Regression.

BiLSTM: simple. The BiLSTM trained on $l \in L_{hl}$ also improved its performance compared to the ones trained on $l \in L_{ll}$: all categories now have an F1-score greater than 50% (see Table 8). Just like its Logistic Regression counterpart before, the BiLSTM excells most for the label *The saviour* with an F1-score of more than 90%. While the simple BiLSTM for the ll labels *Presupposition* and *Authority voice* had similar precision values as the BiLSTM trained on $l \in L_{hl}$, the latter model improved its F1-score for the hl label *The expert* because its recall got better. A similar observation can be made for the hl label *The poet*: accuracy and precision values are comparable to the simple BiLSTM's driving force ll label *Compassion*, but the hl model's recall improved (and with it, its F1-score). The confusion matrix in Table 14 shows that we have relatively many false negatives compared to the model's true predictions, and it generally predicted *The poet* to be absent from paragraphs more than present.

	A	P	R	F1
The saviour	83.62	89.03	92.00	90.49
The expert	63.84	57.63	46.58	51.52
The poet	66.67	78.31	61.32	68.78

Table 8: Accuracy, precision, recall, and F1 score for each higher-level label predicted by a simple BiLSTM.

Looking at Table 9 we can determine that the training on $l \in L_{hl}$ seems to lead to models predicting labels more reliably better than the Random Baseline (except for *The expert* in Logistic regression which is about as good as the Baseline model). The fact that the Logistic Regression model performs best in the category *The saviour* can likely be traced back to its overgeneralisation tendencies. The BiLSTM's model performance for *The expert*, while being better than the Logistic Regression and Random Basline models, does not improve as astonishingly with regards to its percentage points as the Logistic Regression does, but the BiLSTM trained on $l \in L_{ll}$ also predicted better for the relevant ll labels before. It seems that the BiLSTM generally predicts more positive *The expert* occurrences than the Logistic Regression does, both correctly and incorrectly. Finally, the Logistic Regression model outperformed the BiLSTM again in its classification for *The poet*. Looking at the confusion matrices in Table 15 it seems that the BiLSTM generally predicted more pessimistically for this label, leading to 11 more true negatives instead of false positives, but also to more false negatives and fewer true positives.

	Random	LogReg	BiLSTM
The saviour	76.64	**91.14**	90.49
The expert	39.71	39.18	**51.52**
The poet	61.17	**76.15**	68.78

Table 9: Comparing F1 scores for each higher-level label from Random Baseline, Logistic Regression, and BiLSTM.

Overall, it seems that both model types profit from training and predicting on grouped (i.e. hl) rather than individual (i.e. ll) labels.

Table 12 in A illustrates a direct comparison of our results to Pérez Almendros et al. (2020a)'s k-fold cross-validation BiLSTM regarding the performance of ll multi-label classification. We can see that our Random Baseline also dwarfs Pérez Almendros et al.'s BiLSTM regarding the prediction of *The poorer, the merrier*. Other than that, their BiLSTM reaches the highest F1-score for *Presupposition* and *Compassion* while the other six $l \in L_{ll}$ are classified best with one of our four ll-models. And even though our k-fold cross-validated BiLSTM is only the best performing model for *Authority voice* compared to our other models, it still outperforms Pérez Almendros et al.'s BiLSTM in two thirds of the labels (excluding *The poorer, the merrier* where both have nothing to show). The most prominent performance differences can be noted for the labels *Presupposition, Authority voice, Metaphor,* and *Compassion* where both models diverge more than ten percentage points from each other; in the case of *Metaphor* even close to twenty. This degree of differences in performance is surprising since we modelled our cross-validation BiLSTM as closely to Pérez Almendros et al.'s cross-validation BiLSTM architecture as possible. So either we unconciously diverged from their implementation in some crucial respect or their updated dataset really did make a tremendous difference somehow, or a combination of both.

8 Error Analysis & Discussion

Starting from the very beginning, there are a lot of potential problems to be addressed. Firstly, the input data: Because of the low number of datapoints and aggreviated by the unbalanced distribution of categories, the Don't Patronize Me! dataset (Pérez Almendros et al., 2020a) does not provide the best base for confidently learning patterns that refrain from overly simplifying based on or otherwise suffering from the label imbalance. Furthermore, with removing some of the punctuation from the input data we performed a pre-processing step for which we have no information whether (Pérez Almendros et al., 2020a) did the same – thus potentially skewing our model performances and direct comparability.

Our embeddings also provide some food for thought: For our Logistic Regression models, the tf-idf embeddings were learned solely based on the train data. This results in some loss of unknown words when converting the test data into paragraph-level sentence embeddings by means of the 'trained' converter. More daunting, perhaps, is the impact of the Google News word vector embeddings. On the one hand, using these pre-trained word embeddings seems quite obvious because the Don't Patronize Me! dataset is based on newspaper articles just as the Google embeddings. However, we found that the Google News does not seem to recognise British English spelling which is problematic considering the twenty different English speaking countries the Don't Patronize Me! dataset draws news articles from. Those pre-trained word embeddings thus do not encompass tokens like 'modernisation', 'marginalisation' and 'patronise' which could be potentially important tokens for models to learn from. Something which might be less impactful for our purposes than the British English spelling problem but is still worth pointing out is that the Google News embeddings also seem to have trouble with some proper names (e.g. 'DUSIB', 'TheDignityProject', 'Mumassaba') and numbers (e.g. '400,000', '2010').

Turning to our models themselves, while it is astonishing how many correct labels our Logistic Regression models were able to predict by more or less categorically guessing for often-appearing and against seldomly-appearing labels. We might have been able to improve the Logistic Regression models' performance a bit more by trying out different parameters for the arguments `C` and `class_weight`, but the models' blunt overgeneralisation still reinforces that this kind of simplistic model is unfit for classifying more complex problems such as PCL. Both types of BiLSTM models seem to have learned more nuances regarding the classification of $l \in L_{ll}$ and generalise less. Furthermore, a simple BiLSTM model seems to predict for $l \in L_{hl}$ more reliably than for $l \in L_{ll}$. It is still appropriate to question the general architecture we applied for the BiLSTM models. With our hyperparameter selection we intended to increase comparability to Pérez Almendros et al. (2020a)'s BiLSTM model; however, for example a bigger hidden size in the BiLSTM layers might have helped to model to identify certain labels better through higher levels of abstraction. Regarding all models implemented with k-fold cross-validation, it remains

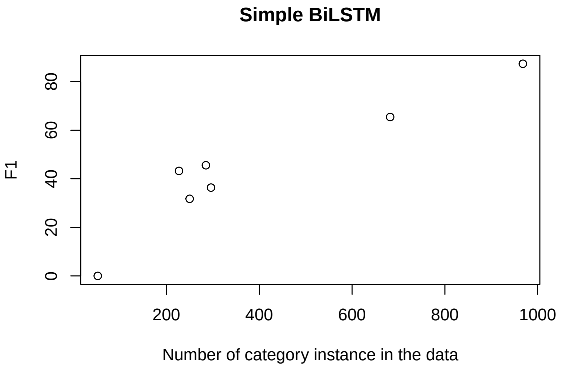

Figure 4: Relation of the amount of lower-level category instances in the data to the simple BiLSTM's F1-score for each category.

Correct predictions	
Paragraph	**Gold**
"He depicts demonstrations by refugees at the border post, their catastrophic living conditions and the desperate attempt of several hundred to cross a river a few kilometres from the camp to get into Macedonia on 14 March 2016."	comp.
"He wants more done now to help those in need."	unba.
"We have the opportunity to give the gift of love, to shine a light in the darkness of despair, to share with others who are in need, to comfort those who are sad or lonely ."	unba., meta., comp.
"As exemplified by the teenager, let us all be kind to one another and help those in need without discriminating against one another based on race or religion."	unba.
"Bond went out of his way to help the less fortunate, often going on the road with Kim to take food to the homeless."	unba., comp., shal.

Table 10: Examples where the k-fold cross-valdiation BiLSTM model predicted lower-level labels identical to the gold labels.

questionable whether the best performance scores on the train/validation set is truly meaningful enough to ensure good model performances on unseen data.

More generally, the scores of $l \in L_{ll}$ seem to correlate with the frequency of their respective label in the dataset (see Figure 4).

Regarding the classification for $l \in L_{hl}$ – it is an interesting result to see *The expert* and *The poet* being more accurately classified than the ll-labels they are comprised of, from which one could infer a pointer strengthening the taxonomy's assumption to group the $l \in L_{ll}$ to hl-labels in this way. However, due to the small amount of underlying data points it remains somewhat unclear whether this improved performance is truly due to the $l \in L_{ll}$ co-occurring this way or whether the increased amount of instances per $l \in L_{hl}$ merely alleviates the poor representation of some $l \in L_{ll}$ in the data. It might have been a good idea to compare the performance of models trained on $l \in L_{ll}$ for hl-labels when inferring $l \in L_{hl}$ from the predicted $l \in L_{ll}$ to the performance of models actually trained on $l \in L_{hl}$ to see if the ll-label grouping truly does have a positive effect on classification performance.

Looking into the k-fold cross-validation BiLSTM predictions for $l \in L_{ll}$ more deeply, we can observe that there are 27 paragraphs for which the predicted labels match the annotated gold labels precisely. Table 10 illustrates that identical labelling often relates to recognising *Unbalanced power relations* either alone (ex. 2,4) or in combination with other categories (ex. 3,5). This trend is not exclusive, however, as illustrated by the first example which has the singular label *Compassion* associated to it. This greater variation also lets our BiLSTM stand out compared to our Random Baseline, which not only had only 6 identical predictions but had mostly statistically prone lucky guesses with only labelling *Unbalanced power relations*.

Reviewing instances of non-identical labelling we find that our model is sometimes wrong regarding one category (ex. 1,2,3 in Table 11), sometimes regarding many labels (ex. 4,5). In several instances the BiLSTM only classified *Unbalanced power relations* (ex. 5); this is probably related to that category's omnipresence. Cases of non-identical labelling also put the gold labels into question: For instance, the first example in Table 11 is annotated to contain only the PCL type *Shallow solution* while our model labels *Presupposition* as second appearing category. For "Many refugees don't want to be resettled anywhere, let alone in the U.S." I would argue that it is justified to assume a *Presupposition* to be present. This shows again that PCL is a very subtle and also individually perceived type of language, begging the question whether the opinion of two to three annotators with a very homogeneous background is truly sufficient here.

9 Ethical Considerations

Bluntly put, the models presented here are not fit to be employed for any application out in the wild. The kind of context and level of abstraction needed for this task by far exceed what our models can handle. Furthermore, from what results our experiments yielded, applying k-fold cross-validation did not justify the amount of energy resources it took to train models with it. While we chose to investigate the effect on model performance when training on lower- and higher-level labels, it is important to put more effort into examining potential subgroups in the data. For example, different English speaking countries could have different ways of expressing PCL or differ in what is perceived as PCL contingent on their culture. This kind of cultural contingency obviously also impacts the transferability to even more different cultures and other languages. But speaking of transferability, how promising is the approach of PCL classification taken here for real-life applications?

Research about PCL so far has mostly focused on direct speech (e.g. Brown and Draper (2003), Fox and Giles (1996), Keaton and Giles (2016)); approaches in NLP more often focused on less subtle forms of hurtful language use (e.g. Mendelsohn et al. (2020), Zampieri

Incorrect predictions		
Paragraph	Gold	Pred
"Many refugees don't want to be resettled anywhere, let alone in the US."	shal.	shal., pres.
"Real poverty of Britain: Shocking images of UK in the Sixties where poor really meant poor. THESE hard-hitting photographs offer a glimpse into the harrowing day-to-day for poor families living in Britain during the Sixties."	auth., comp.	meta., comp.
"A top health official said today that the government could consider subsidies to help poor families pay for healthy food – or imposing taxes on unhealthy products – if other efforts fail to encourage better eating habits among Hong Kong residents. Dr Regina Ching from the Health Department said such moves could be explored as a way to cut levels of chronic illnesses in the city, such..."	unba., pres.	unba., comp.
"Stephanie envisioned a model whereby women in need would be taught a skill such as handcrafts , and be paid a fair trade wage for their efforts. In turn, the women who had experienced hardship and marginalisation would grow to become self-sufficient. With the products then sold in Australia via an online store and a network of retailers, the profits would be reinvested in Seven Women to continue the cycle of empowerment."	unba., shal., pres., comp.	unba., pres., auth.
"Mari?tte Coetzee from Stofberg Family Vineyards (whose Mia Chenin Blanc 2016 was the garagiste trophy winner at last year's Michelangelo Awards and the recipient of four Platter's stars for the Mari?tte Chenin Blanc 2016), says: 'We can be extremely proud of the current women winemakers in our industry, especially considering most of them are juggling a family along with the long working hours."	unba., comp., tptm.	unba.

Table 11: Examples where the k-fold cross-valdiation BiLSTM model predicted lower-level labels not identical to the gold labels.

et al. (2019), Basile et al. (2019), Sap et al. (2019)) or dangerous communication practices (e.g. Conroy et al. (2015)) in direct or indirect communication, or dealt with subtle PCL in direct communication settings (Wang and Potts, 2019). All of these approaches are arguably very different from Pérez Almendros et al. (2020a) which neither deals with direct communication nor with more explicitly harmful language. Potential applications for detecting and classifying PCL in such an indirect media setting such as Pérez Almendros et al. (2020a)'s newspaper articles thus come with several challenges: The style of language use in newspapers is very different compared to any form of direct communication, independent of whether communication takes place synchronously or asynchronously. Thus, it would be hard to draw from findings gained in other communication settings even though they might draw a clearer picture of what can constitute PCL. This difficulty in detection processes is exacerbated by the social impact media has globally – any PCL detection application would be required to classify maximally reliably in order for it to not falsely accuse PCL use but also being reliably confident in classifying something as PCL so as to actually reduce harmful language in the world. Also, there are so many more vulnerable groups in the world than what Pérez Almendros et al. (2020a) investigate, e.g. ethnical minorities like the Sámi, children, people with a chronical illnes, etc. Would a model trained with a limited set of vulnerable groups even be able to detect and classify all PCL uses well? Steps would be needed to ensure such an application would not just jump to conclusions based on pre-set keywords.

In any case, due to the different language use in newspaper articles, detection models trained on such articles would never be able to be applied to any direct communication, be it face-to-face or by means of (a)synchronous media. Still, it is desirable to be able to detect PCL in 'scripted' language uses, not just newspaper articles but also for example news reports or documentaries, because this type of media has a huge impact on day-to-day life, being targeted at everyone while never giving vulnerable groups the possibility to challenge occurring PCL. Unchallenged, this PCL can further feed stereotypes (Fiske, 1993), fuel discriminatory behaviour (Mendelsohn et al., 2020) and strengthen power-knowledge relationships (Foucault, 1980). Perhaps, if the desired degree of detection reliability is impossible to achieve, it would already constitute an improvement to issue a warning stating that the current phrasing might be condescending/patronising and thus still raise awareness to the PCL producer and/or recipients.

10 Conclusion

In this project we set out to compare the performance of two differently complex models for the task of multilabel PCL classification for two levels of abstraction. For the lower-level label prediction, our simple model, the Logistic Regression, achieves quite astounding accuracy scores but it overgeneralises too much on the underlying label distribution to be safely put to use in an applied setting. Our more complex model, the BiLSTM, performs much more reliably than the Logistic Regression model. It achieves higher scores for the prediction of most labels and is less prone to overgeneralisation. Any application choosing to be based on a BiLSTM for this task should still take great care and not take its results at face value. The training cost for implementing a k-fold cross-validation for any of our models did not pay off regarding result improvements.

The classification of higher-level labels works much more reliably for both the Logistic Regression and the BiLSTM. It seems worthwhile to investigate the performance of higher-level labels more thoroughly in the future. It is difficult to make any strong, confident claims about these results, however, when most results/trends could be side effects of the dataset being small and/or the label distributions being unbalanced.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Angie Brown and Peter Draper. 2003. Accommodative speech and terms of endearment: Elements of a language mode often experienced by older adults. *Journal of Advanced Nursing*, 41(1):15–21.

Francois Chollet et al. 2015. Keras.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

Mark Davies. 2013. Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. Available online at https://corpus.byu.edu/now/.

Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American psychologist*, 48(6):621.

Michel Foucault. 1980. *Power/knowledge: Selected interviews and other writings, 1972-1977*. Vintage.

Susan Anne Fox and Howard Giles. 1996. Interability communication: Evaluating patronizing encounters. *Journal of Language and Social Psychology*, 15(3):265–290.

Shaughan A Keaton and Howard Giles. 2016. Subjective health: The roles of communication, language, aging, stereotypes, and culture. *International Journal of Society, Culture & Language*, 4(2):1–10.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Sik Hung Ng. 2007. Language-based discrimination: Blatant and subtle forms. *Journal of Language and Social Psychology*, 26(2):106–122.

David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Carla Pérez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020a. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carla Pérez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020b. Don't patronize me! data statement. Version 1.1.

Carla Pérez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2021. Semeval 2022 task 4: Patronizing and condescending language detection.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

P. Szymański and T. Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

A Additional Visualisations

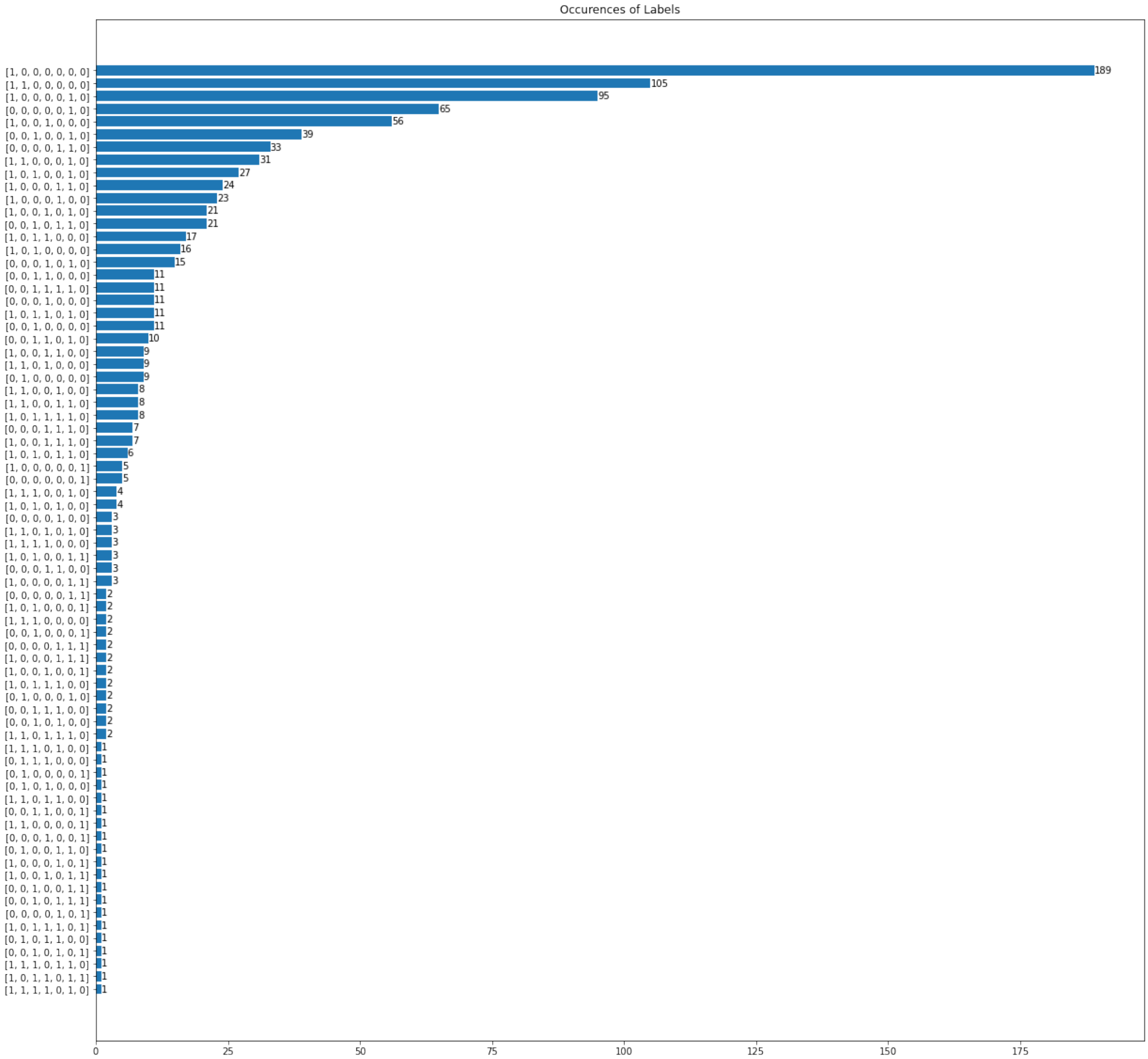

Figure 5: Distribution of lower-level label sets (gold) in the train data.

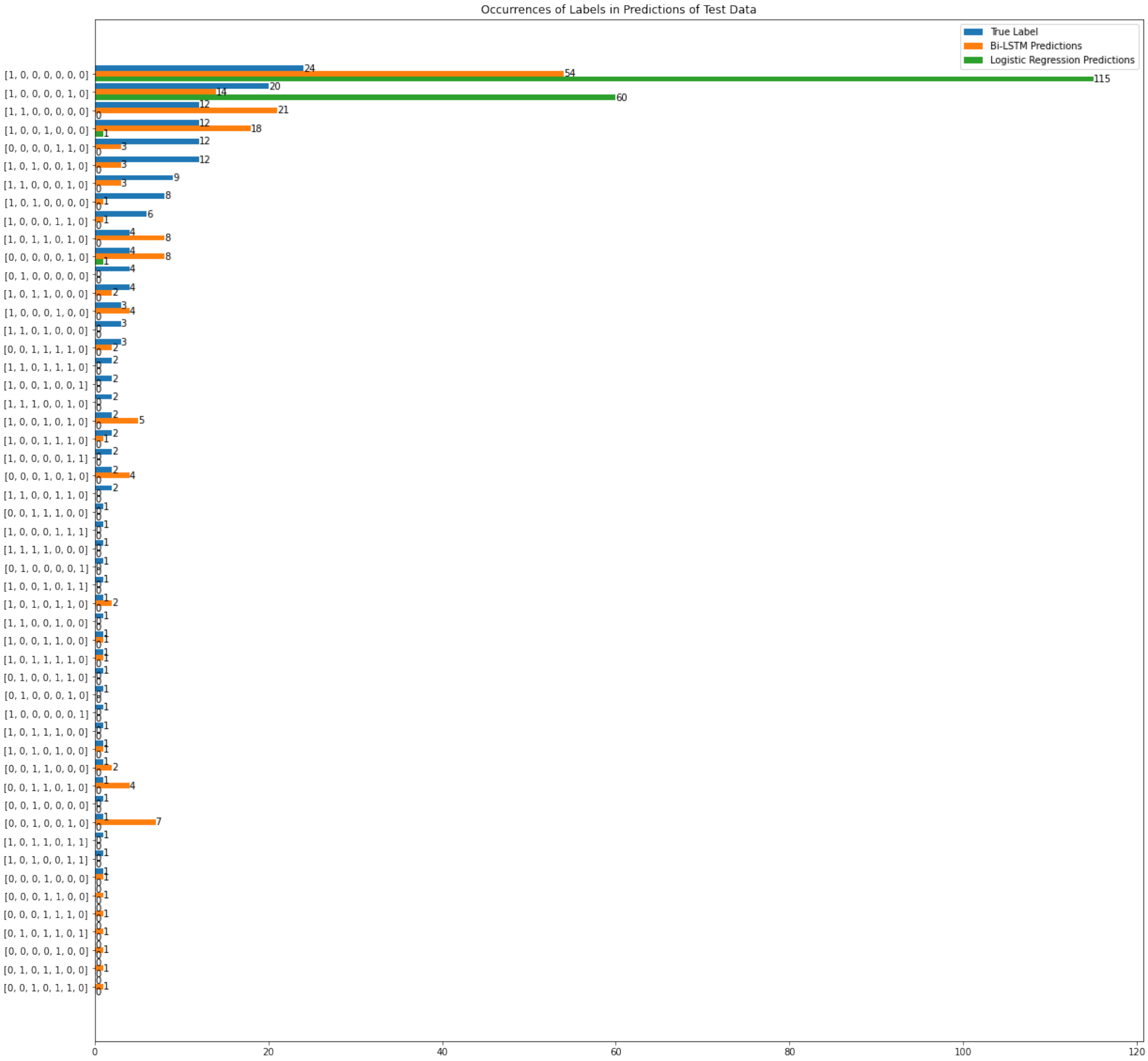

Figure 6: Distribution of lower-level label sets (gold and predicted) in the test data.

				Random			BiLSTM kCV					
				P	R	F1	P	R	F1			
Unb. power rel.				82.5	69.23	75.29	83.94	84.58	83.92			
Shallow solu.				25.0	23.08	24.0	64.06	31.84	40.46			
Presupposition				38.64	37.78	38.2	52.44	36.02	**41.15**			
Authority voice				22.86	17.39	19.75	42.54	21.27	25.71			
Metaphor				30.56	28.81	29.33	7.83	1.99	3.14			
Compassion				52.56	43.62	47.67	74.48	69.86	**71.34**			
The p., the mer.				16.67	10.0	**12.5**	0.00	0.00	0.00			
	LogReg			LogReg kCV			BiLSTM			BiLSTM kCV		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Unb. power rel.	80.68	99.30	**89.03**	80.68	99.30	**89.03**	90.30	84.62	87.36	90.00	81.82	85.71
Shallow solu.	0.0	0.0	0.0	0.0	0.0	0.0	45.71	41.03	**43.24**	48.39	38.46	42.86
Presupposition	0.0	0.0	0.0	0.0	0.0	0.0	57.14	26.67	36.36	41.67	22.22	28.99
Authority voice	100	2.17	4.26	100	2.17	4.26	54.55	39.13	45.57	46.67	45.65	**46.15**
Metaphor	0.0	0.0	0.0	0.0	0.0	0.0	41.67	25.64	**31.75**	40.00	25.64	31.25
Compassion	78.69	51.06	61.04	80.0	46.81	59.06	77.94	56.38	65.43	63.41	55.32	59.09
The p., the mer.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 12: Results for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem using lower level labels. The upper part presents results from our random baseline and the 10-fold cross-validation BiLSTM from Pérez Almendros et al. (2020a) to serve as comparison to our model results in the lower part.

	Random			LogReg			BiLSTM		
	P	R	F1	P	R	F1	P	R	F1
The saviour	84.68	70.00	76.64	84.66	99.33	**91.41**	89.03	92.00	90.49
The expert	42.87	36.99	39.71	79.17	26.03	39.18	57.63	46.58	**51.52**
The poet	63.00	59.43	61.17	74.11	78.30	**76.15**	78.31	61.32	68.78

Table 13: Results for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem using higher level labels.

Labels	Random		LogReg		LogReg kCV		BiLSTM		BiLSTM kCV	
1	13	21	0	34	0	34	21	13	21	13
	44	99	1	142	1	142	22	121	26	117
2	111	27	138	0	138	0	119	19	122	16
	30	9	45	0	39	0	23	16	24	15
3	105	27	132	0	132	0	123	9	118	14
	28	17	45	0	45	0	33	12	35	10
4	104	27	131	0	131	0	116	15	107	24
	38	8	45	1	39	0	28	18	29	21
5	113	25	138	0	138	0	124	14	123	15
	28	11	39	0	39	0	29	10	29	10
6	46	37	70	13	72	11	68	15	53	30
	53	41	46	48	50	44	41	53	42	52
7	162	5	167	0	167	0	165	2	165	2
	9	1	10	0	10	0	10	0	10	0

Table 14: Confusion matrices of our model's performances regarding each lower level label. 1- Unbalanced power relations; 2- Shallow solution; 3- Presupposition; 4- Authority voice; 5- Metaphor; 6- Compassion; 7- The poorer, the merrier. Each confusion matrix has the following layout: upper row: true negative, false positive; lower row: false negative, true positive.

Labels	Random		LogReg		BiLSTM	
1	8	19	0	27	10	17
	45	105	1	149	12	138
2	68	36	99	5	79	25
	46	27	54	19	39	34
3	34	37	42	29	53	18
	43	63	23	83	41	65

Table 15: Confusion matrices of our model's performances regarding each higher level label. 1- The saviour; 2- The expert; 3- The poet. Each confusion matrix has the following layout: upper row: true negative, false positive; lower row: false negative, true positive.

B Personal Reflection

Overall, I feel like this project went rather well. We ended up not implementing BERT because that would have been too time-consuming and we also had to cut short on further ideas and additional explorations that we would have liked to do in order to be able to finish up well what we decided to implement. It was hard to decide to not follow up on things that we felt would have enriched our project further, but I believe it was a valuable lesson to learn to set sensible limitations to one's projects in order for it to not get out of hand. I definitely gained confidence in working on projects in this field, and I learned how to use git and what different approaches to coding together can be. Whether or not my application of everything related to this project conforms to standards in this field I leave to you to judge.

I think it is safe to say that I did my best in being a good group member – just like the rest of my group. We met regularly to share our progress and discuss about it, kept self-set deadlines and generally invested a lot of time and effort into this project. Antonia and I tried out pair programming and are jointly responsible for everything BiLSTM-related, including the BiLSTM part in `0. Data Preprocessing`, the `embedding_functions`, and `2. Bi-LSTM`. Furthermore, I implemented our idea for the random baseline `(3.)` and did the error analysis part in `4. Analysis`. I also created several tables for our reports. In the end we did not get to investigate errors and their implications (linguistically) to the extend that I would have liked because many more subtle things came up for discussion only when we were writing our individual papers and were thus unable to add any coding.

My main take-away from this project is probably a) that it is important to set clear goals and limits so that projects are sound while also not getting out of hand, b) that the process of making code presentable to others takes much longer than I thought, and c) that one should have more intense intermediate summaries and in-depth discussions about them in case the code-deadline is before the paper deadline because otherwise you notice things while writing that you then cannot follow up on afterwards.