Automated Multi-label Classification of Patronizing and Condescending Language (PCL) in English News Articles

Lara Pfennigschmidt (791642)
Final Project Report
University of Potsdam / Advanced Natural Language Processing 2021/2022
pfennigschmi@uni-potsdam.de

Abstract

The detection and classification of Patronzing and Condescending Language (PCL) is important to reduce negative effects and bias of often well-intentioned statements. Resulting from this insight, this year's SemEval Challenge includes a task about PCL detection and classification and provides the "Don't Patronize Me!" dataset along with a taxonomy for PCL categories. This paper intends to go about this task and explain our work for data preprocessing, training of logistic regression and bidirectional LSTM models with and without cross-validation, as well as an extension of the task by predicting PCL category labels on different levels of the taxonomy. The BiLSTM models showed better results on the original task, whereas the logistic regression provides more time-efficient predictions and is more effective for predicting the higher-level labels. Also, we provide a random baseline to which the results can be compared with.

1 Introduction

Patronizing and Condescending Language (PCL) as a phenomenon of speech and written text can be observed around the world. Although its existence can be subtle and difficult to detect, it is an important matter to be aware of. The use of PCL evidently leads to bias and discrimination, often towards already vulnerable groups. (Nolan and Mikami, 2013). To reduce negative effects of normally well-intended utterances, it is vital to detect PCL and helpful to categorize it, for spreading awareness and providing explanations of why an expression contains PCL and which kind. For this report, we investigated whether automated language models can pick up the subtle cues of PCL and are able to label them with several categories.

2 Related Work

Our work directly relates to Pérez Almendros et al. (2020a), who have published the "Don't Patronize Me!" dataset, a curated and annotated dataset of English news articles containing PCL instances. The news articles were sampled from the News on Web (NoW) corpus by Davies (2013). Additionally, they proposed a taxonomy of the different categories of PCL splitting three higher-level labels into seven lower-level labels as follows:

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

They already experimented with some models and showed, that the most promising results came from BERT-models.

PCL classification was also part of 2022's SemEval Challenge: "Task 4: Patronizing and Condescending Language Detection, Subtask 2: Multi-label classification. Given a paragraph, a system must identify which PCL categories express the condescension." (Pérez Almendros et al., 2021). This task is also based on the previously mentioned "Don't Patronize Me!" dataset.

Apart from this effort, not much else has been done in the field of automated PCL detection.

Our work is mainly concerned with the task of the SemEval Challenge. In addition, it extends this task by providing comparisons between two different language models trained for the original task and for predicting the three higher-level labels of the taxonomy in contrast to the seven lower ones.

3 Problem Statement

The task at hand was a multi-label classification of paragraphs: given a paragraph of an English news article, a language model should predict which of the seven labels of PCL as proposed before apply to the text. Take the sentence from Quote 1 as an example. This statement expresses Patronizing and Condescending Language towards immigrants. It does so by raising "Compassion" through describing the situation in an exaggerated way, 'too dirty, too cold and too lonely [...]', so that readers

> Sheepherding in America has always been an immigrant 's job , too dirty , too cold and too lonely for anyone with options .

Figure 1: paragraph ID: 773,
article ID: @@1759840,
Pérez Almendros et al. (2020a)

pity immigrants for the hard work they supposedly have to do. It also exerts an example of "Unbalanced Power Relations", dividing the people into 'us' who are not sheepherding and 'them', who are. The last label this paragraph was tagged with is "Presupposition": there is no real proof for sheepherding only being an immigrant's job or it always being cold and lonely or only a job if you have no other options. This does not only patronize immigrants, but sheepherders in general. With this example, we can get an idea of why this classification is important.

Based on the taxonomy, we extended this task by training models to predict the three higher-level labels as well. Also, we provide a newly proposed random baseline for this problem that accounts for statistical variations in the occurring categories.

4 Data

4.1 Description

The dataset we used was the training portion of the "Don't Patronize Me!" dataset by Pérez-Almendros et al. that was also provided to the SemEval Challenge 2022 Task. The data contains paragraphs from English news articles, labeled with the PCL categories they express. Additionally, each instance of PCL is limited to a subsection of the paragraph, if applicable, expressing where exactly the condescension lies. The paragraphs are sampled from 20 different English speaking countries and apply to 10 vulnerable groups: disabled, homeless, hopeless, immigrant, in-need, migrant, poor families, refugees, vulnerable, and women. The languages are close to being uniformly distributed, but the distribution across vulnerable groups varies (see 2). The resulting labels arise very differently, as can be seen in 3 [1].

4.2 Statement

The data statement (Pérez Almendros et al., 2020b) states that three female annotators aged between 25 and 35 had annotated this dataset. Their first language is Spanish but they are bilingual in English. Two of them annotated the complete dataset while the third resolved disagreements. Because the annotators are an all female group, they might have been biased towards patronizing language use against women.

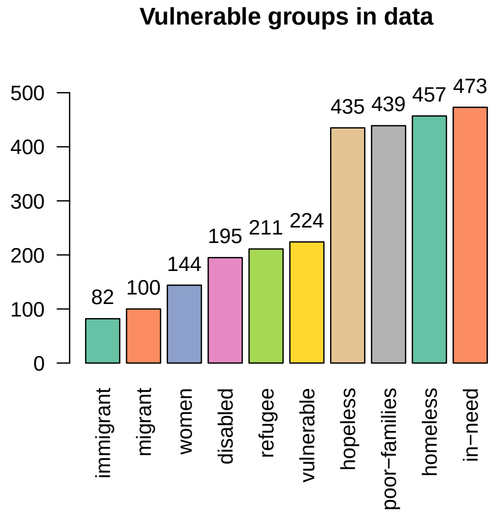

Figure 2: Distribution of vulnerable groups in paragraphs containing instances of PCL.

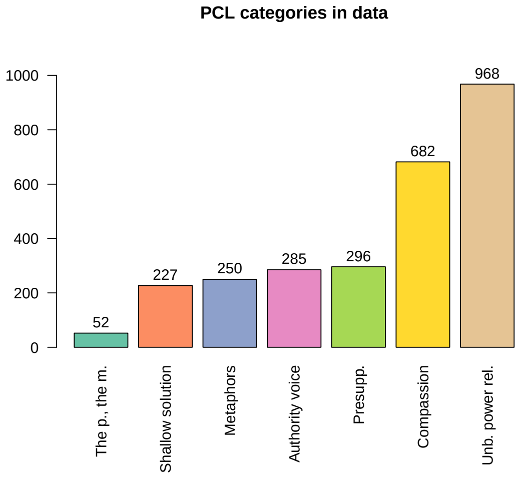

Figure 3: Distribution of types of PCL in paragraphs containing instances of PCL.

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

4.3 Preprocessing

For training purposes, the paragraphs had to be converted into paragraph embeddings and word embeddings for different models. The labels were easily converted into multi-hot-encoded vectors and the higher-level labels could be inferred from the lower-level labels in the way the taxonomy states. We removed one paragraph from the dataset as it was an outlier in its lengthiness. Because we only had the training portion of the complete "Don’t Patronize Me!" dataset, we introduced our own train-test split to the data (following mentions of train and test data refer to this split). The split has been done 80/20 and was stratified by the amount of paragraphs each PCL category has been labeled with, using Iterative Stratification from scikit-multilearn (Szymański and Kajdanowicz, 2017) as is recommended for multi-labeled data (random_state = 1).

5 Experiments

Our experiments include two different language models: logistic regression and bidirectional Long-Short-Term-Memory (BiLSTM) models. Also we created our own random baseline for our specific train-test-split. The first experiment is prediction in terms of the original task. The second experiment uses 10-fold cross-validation for finding a best model on a validation set. Our third experiment is a comparison between the predictions for the lower-level labels and the higher-level labels. The idea for training on the higher-level labels is to have fewer labels and more samples per label. If the taxonomy is sensible, this should be a valid option for PCL classification as well.

5.1 Models

5.1.1 Random Baseline

Our random baseline calculates the percentage of training paragraphs appearing in each category, and predicts test paragraphs belonging to a category with the same probability. Because our train-test-split is stratified by category, this assumption is more fitting than a 50/50 prediction of a label. The random number generator was seeded with the number 11, floating point numbers equal or above the probability were classified as 0 whereas numbers below the probability were classified as 1.

5.1.2 Logistic Regression

For the logistic regression, we used scikit-learn’s LogisticRegression algorithm with default parameters and a MultiOutputClassifier-Wrapper to handle multiple labels. This internally calculates a logistic regression for each category separately and combines the outputs into a single multi-label output. We also compared this to a plain logistic regression that predicts the complete label as a string (random_state = 1) The paragraphs were converted to embeddings (one embedding per paragraph) using scikit-learn’s TfidfVectorizer. The test data was converted to paragraph embeddings using the same "trained" converter, which resulted in a loss of unknown words in the test data. For the prediction of higher and lower levels, only the train input needed to be replaced to represent the correct set of labels. The 10-fold cross-validation has been done with scikit-learns KFold algorithm (shuffle = true, random_state = 1).

5.1.3 Bidirectional LSTM

We used the functional API of Keras Chollet et al. (2015) for creating and visualizing our BiLSTM model. The model reproduces the structure of the BiLSTM from the paper (Pérez Almendros et al., 2020a) to be comparable. We took provided hyperparameters for:

- number of layers: 2 (C. Pérez-Almendros, personal communication, March 17, 2022)
- number of units per layer/hidden size: 20
- number of epochs: 300, with early stopping and a patience of 10
- optimizer: Adam
- drop-out rate: 0.25%

Other set parameters include a batch size of 32, and maximum length of input sequences of 191. The model architecture is visualized in Fig. 4.

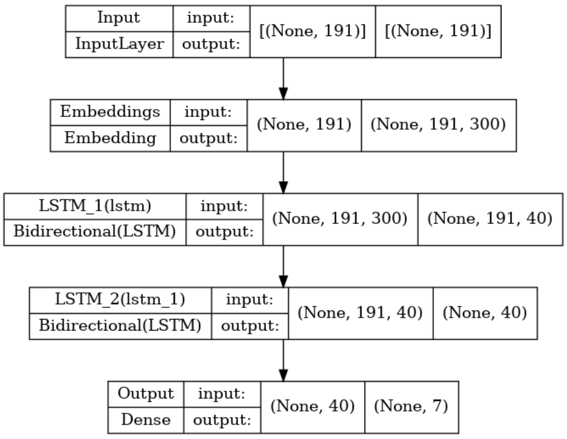
Input | input: | [(None, 191)] | [(None, 191)]
InputLayer | output:
Embeddings | input: | (None, 191) | (None, 191, 300)
Embedding | output:
LSTM_1(lstm) | input: | (None, 191, 300) | (None, 191, 40)
Bidirectional(LSTM) | output:
LSTM_2(lstm_1) | input: | (None, 191, 40) | (None, 40)
Bidirectional(LSTM) | output:
Output | input: | (None, 40) | (None, 7)
Dense | output:

Figure 4: The structure of the BiLSTM, decomposed into its’ layers.

Because the BiLSTM takes sequences of words as input, only the words needed to be converted into embeddings and not whole paragraphs. The paragraphs are tokenized while some more punctuation (-, ., /) is replaced by white-space and used to split up words even further. The embeddings were created using the pre-trained Google News embeddings (Mikolov et al., 2013). Words unknown to the embedding matrix were left out. Word embedding sequences were then padded if necessary with empty words up to a sequence size of 191. The loss function used for training was binary cross-entropy loss. The output predictions of the model were probabilities, so we had to make them discrete manually by splitting at 0.5 into labels of 0 and 1. The 10-fold cross-validation for the BiLSTM uses the same function that was used for the logistic regression (random_state = 11).

For predicting the lower and higher-level labels, we modified the output size of the output layer to return seven or three labels respectively, as well as replaced the training input with the corresponding labels.

5.2 Evaluation

For the evaluation and selection of our models, we mainly used the average F1 score as well as the F1 score for each PCL label separately. Further gathered metrics include accuracy, precision and recall, as well as confusion matrices per category (Table 13 and Table 12). The following sections show the results of our experiments for each model.

5.2.1 Lower-level labels

Logistic Regression A plain logistic regression that predicts the complete label as a string scores about 0.44 accuracy on the set it trained on. The score on the test data drops down to 0.14 accuracy. The multi-label logistic regression that handles each label individually scores a bit lower on the train data with a mean accuracy of 0.30, but comes in higher for the test data with 0.175 accuracy. Seeing that the accuracy scores on the training data are already low, it does not surprise that it drops further when predicting the test data. But the mean accuracy is a harsh metric for a multi-label classification because each label has to be set correctly for each sample, and this is rather unlikely (Pedregosa et al., 2011). Therefor, we calculated the metrics per label and show them in Table 1.

	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.

We can see that the model fails to predict a single positive example for four of the seven categories, namely "Shallow Solution", "Presupposition", "Metaphor", and "The poorer, the merrier". Looking at the accuracy per category, we can see that the model still gets high scores, above 0.65, which suggests that the model has learned to predict these categories to never occur based on their statistical minority. This can also be seen in the confusion matrices per label as most samples are true negatives (Table 13). Something very similar is happening to the first label of "Unbalanced Power Relations". Here the model almost always predicts this label for each paragraph, because it is statistically highly likely. For "Authority voice", only one single example is classified as positive, which results in a precision measure of 1, but recall and therefor F1 measure are lacking. At least for "Compassion" the model predicts both positive and negative samples, and it also seems to do better than predicting purely by chance (Table 3).

BiLSTM The bidirectional LSTM scores 0.64 on the train data, which is much higher than what the logistic regression achieved. The metrics per label can be found in 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.

We can see that the BiLSTM also has a problem with predicting "The poorer, the merrier" as does the logistic regression. But the overall scores are higher, importantly the F1 scores. For "Unbalanced Power Relations" and "Compassion", the F1 scores reach quite high (above 0.6), whereas for the other 4 labels the score is lower than 0.5. This suggests that the BiLSTM has picked up on more cues to classify PCL, but looking at the numbers, it is still not a sufficiently good model for this task.

Comparison Concluding this experiment, we will show the F1 scores per label for the random baseline compared to the logistic regression and BiLSTM results in Table 3. The BiLSTM scores highest for four labels, the logistic regression only for the first, and for the remaining two categories, our random baseline provides better results than our trained models.

	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.

5.2.2 k-Fold cross-validation

With the 10-fold cross-validation, we trained the models on 10 different folds of the training data and kept part of the data as a validation set. Then, we selected the model with the best mean F1 score over all labels, although there were models with better validation accuracy in some folds.

Logistic Regression For the logistic regression, the predictions of the selected model are exactly the same as in the previous experiment, except for "Compassion", which got a higher accuracy, but lower recall and slightly lower F1 measure as can be seen in Table 4.

	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, chosen via 10-fold cross-validation.

Because the cross-validation did not improve our model, we can conclude that it is not helpful for logistic regression in this case. This may still have different reasons, either it does not harmonize well with logistic regression in general or the underlying distribution of the data is the culprit (more on this in section 5.3.3).

BiLSTM The cross-validated BiLSTM improves the F1 score for one of the seven labels ("Authority voice"), but for the others it remains at a similar level ("Unbalanced Power Relations", "Shallow Solution", "Metaphor", "The poorer, the merrier) or drops ("Presupposition" and "Compassion").

	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.

Comparison To conclude this experiment, we can view the F1 scores of the random baseline and the original models plus their cross-validated versions. We can see that the only improved model is the BiLSTM, but the improvements are very slight and cannot reach a higher average than the original BiLSTM.

	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.

These results lead to the conclusion that k-fold cross-validation is not suitable for this dataset, because it only improved the models slightly, although it is recommended state-of-the-art. This may be due to the small size of the dataset to begin with, or because of an unfavorable train-test-split.

5.2.3 Higher-level labels

Logistic Regression For predicting the three higher-level labels, the plain logistic regression reaches higher scores on the train set with 0.75, but drops significantly when tested, only scoring 0.26. The multi-label logistic regression reaches 0.64 on its train data and at least 0.4 on the test data. These are overall better results than the predictions on the lower level could provide.

	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 multi-label logistic regression.

The label of "The saviour" is still labeled almost always (with one exception, Table 12), but the score improves because even fewer negative samples are now present. The model also seems to be more confident in classifying examples as containing "The expert", which improves the overall score. But recall and F1 measure for this category are still behind. "The poet" previously contained two lower labels that were always predicted negatively; now accuracy and precision dropped a little as opposed to only paragraphs with "Compassion", but recall and F1 increased. Overall, this is an improvement for this category.

BiLSTM The BiLSTM scores quite high during training with 0.86, and also the F1 scores per label reach new heights (see Table 8).

	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 BiLSTM.

Comparison The average F1 score of the logistic regression (68.9) is on par with the score of the BiLSTM (70) for this classification task. In Table 9 it can be seen that the logistic regression achieves higher scores for the labels "The saviour" and "The poet" whereas the BiLSTM scores highest for "The expert". Both models provide better or almost as good predictions for the higher-level labels as does the random baseline.

	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.

With these results, predicting the higher labels instead of the lower-level labels does improve overall model performance and should be noted for further research in this field. This also confirms the taxonomy, because the labels are easier to predict if combined.

5.3 Error Analysis

5.3.1 Domain

For the detection and categorization of PCL much context knowledge is needed. World knowledge about the situations of vulnerable groups, concepts used for metaphors, and the effectiveness of actions is required to recognize specific labels, particularly "Metaphor" and "Shallow solution". Added to that, language use is very diverse so not every instance of PCL is detectable.

5.3.2 Data

The data itself is the major source for errors for this task. There are too few samples for some labels, especially for "The poorer, the merrier", so that the models have difficulty picking up on them. The data is not well distributed overall, and there may be subgroup trends influencing the general situation further. The unfortunate distribution is best experienced with the logistic regression, which fails to predict positive samples for four labels. The accuracy of the model is still high, but only because there are so many true negatives. So it overgeneralizes that this label is better not applied, but this also means that it could not infer real information for these labels based purely on the words used.

5.3.3 Preprocessing

During preprocessing the data, we noticed that the Google News Embeddings (Mikolov et al., 2013) mostly include American English spelling, but do not account for British or other versions of spelling. This is problematic, as the news articles were taken from 20 different English speaking countries, who do not all employ American English spelling. Also, we had to remove further punctuation as hyphenated words were not recognized. Some proper nouns could not be mapped either. Our way of dealing with unknown words by simply not accounting for them is also a potential source of error. To counteract the varied distribution of labels we used a stratified split to allocate the categories evenly to train and test set. A later analysis showed, that the sum of categories had been arranged accordingly, but the numbers of each label set now differed significantly (see 5 and 6). The difference in train and test set could explain a lower performance of the models.

5.3.4 Experiments & Models

We went with the default parameters for the logistic regression, which possibly lead to the over-generalization of categories. The class weights could have balanced out the unequal distribution of labels as well. For the k-fold cross-validation, we used the standard implementation of scikit-learn, which is not stratified. The stratified version cannot handle multi-labeled data and was therefore unsuitable. But we could have used the IterativeStratification-method (Szymański and Kajdanowicz, 2017) to emulate our own stratified k-fold cross-validation. But seeing that the label sets then are arranged variably, this could also have led to more errors. At last, our results cannot be compared directly to the original paper's results, because they have worked with a previous version of the dataset according to their account (C. Pérez-Almendros, personal communication, March 16, 2022).

As an example, we show correctly predicted paragraphs in Table 10 from the cross-validated BiLSTM and wronly predicted paragraphs of the same model in Table 11. The correct labels mainly contain "Unbalanced Power Relations" as is inherent to the data distribution, but also other single labeled categories and more complicated multi-labels were predicted correctly. For the wrongly classified example paragraphs, we can see missing labels from the prediction, additional labels that did not occur in the gold label or the count of labels being correct, but not the kind of labels that were predicted.

6 Ethical Considerations

PCL classification is a task that should be automated to explain why utterances contain patronizing statements and to raise awareness for the different kinds. It can help to avoid discrimination and further bias towards vulnerable groups and make people empathize with their counterpart in a conversation. A good field for this ap-

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.

plication is indeed in the area of news articles, because these should be objective and should not increase already present tendencies in condescension. Only the focus on specific keywords and only a handful of vulnerable groups is problematic. Many other people can be patronized who do not fall into the preselected vulnerable groups like people with mental illnesses, elderly and children. Also, this tool should not be used for censorship or to discredit the news, but rather as a hint to writers to overthink their wording in favor of the people being written about. With this task fully automated, one also has to be careful with misclassification or wrongly detected instances of PCL. The results should be handled with care, as the embeddings used seem to prefer American spelling and thus may inherently possess bias towards certain groups of people already. Also, PCL might express itself differently in other areas of language use like social media or spoken language, not forgetting to mention other languages or other cultures as well. To make matters harder, defining and determining what classifies as PCL and what does not is a difficult task in itself and can lead to disagreement even among human annotators. The meaning of an utterance not only depends on the words used but also on the intended intonation. Hence, a statement can be interpreted differently than how it was meant easily, because of missing context information like intonation, gestures, or body language. All in all, PCL detection and classification is a critical task that needs to be solved with care, a lot of thought, empathy, and respect.

7 Conclusion

To conclude our work, we can state the BiLSTM generally seems more fit for the task than the logistic regression. The logistic regression failed to predict four categories of the seven and thus is deemed unsuitable. But it achieved surprisingly good results on the higher-level label prediction and overtook the BiLSTM by scoring higher for two of the three categories. Generally, the higher-level labels were better to predict than the lower-level labels, which might hint at rephrasing the classification task in the future. Unfortunately, the cross-validation did not improve our models and the models are prone to over-generalization, hence accuracy is definitely not the measure of choice to optimize for this task. Summarized, we can state that there is more research needed in the area of automated PCL detection and classification to solve this task responsibly and further experiments should try to compensate for the missing world knowledge that is required for the successful prediction of PCL labels.

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.

References

Francois Chollet et al. 2015. Keras.

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/.

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.

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.

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

A Appendix

A.1 Contribution

A.1.1 Learning Objectives

Our learning objectives were to learn more about PCL, additional language models like BERT and automated parameter tuning, as well as generally applying our gained knowledge to real world datasets and gain even more experience by working on this task. We succeeded on applying our knowledge to this dataset in being able to preprocess it and critically assess our approach with the language models. Because of time constraints and not having a readily available GPU at hand, we decided early on to leave out the BERT-model and to focus more on the other two models and therefor include the analysis of higher vs. lower-label prediction. We also wanted to compare our results to the paper's results, but learning that we have been working with a different version of the dataset than the authors of the paper had, this comparison would not have shown anything reasonable. My personal goals were to work reproducibly and as scientific as possible, because I also wanted to be able to explain errors in the prediction more than just guessing them. The last part fell shorter than I expected due to having enough other content already and not having enough time to dive deeper.

A.1.2 Personal Contribution

With my background in computer science and experience in coding group projects, I was able to plan ahead, introduce my teammates to git, the handling of online repositories, and to coding conventions. I personally worked on the extended scripts for data loading and evaluation, as well as the train-test split of the data. My main part was writing a notebook and exploring the challenges concerned with logistic regression. I also analyzed and visualized the distribution of the train-test-split in the end.

A.1.3 Take-away messages

My main learnings include:

- First programming and then writing the paper does not necessarily work because while writing you notice too many things that are missing or wrong or could have been done better in other ways.
- The nature and quality of the data have a great influence over the results, so much work and thought should go into preprocessing and cleaning, even more than we put in.
- Programming the models correctly and sensibly is not as easy as I made it out to be. When being able to use pre-implemented versions of algorithms, more attention and consideration has to be paid to the parameters and more sets of different parameters should be tried out first, before settling too fast on one approach.

A.2 Tables & Figures

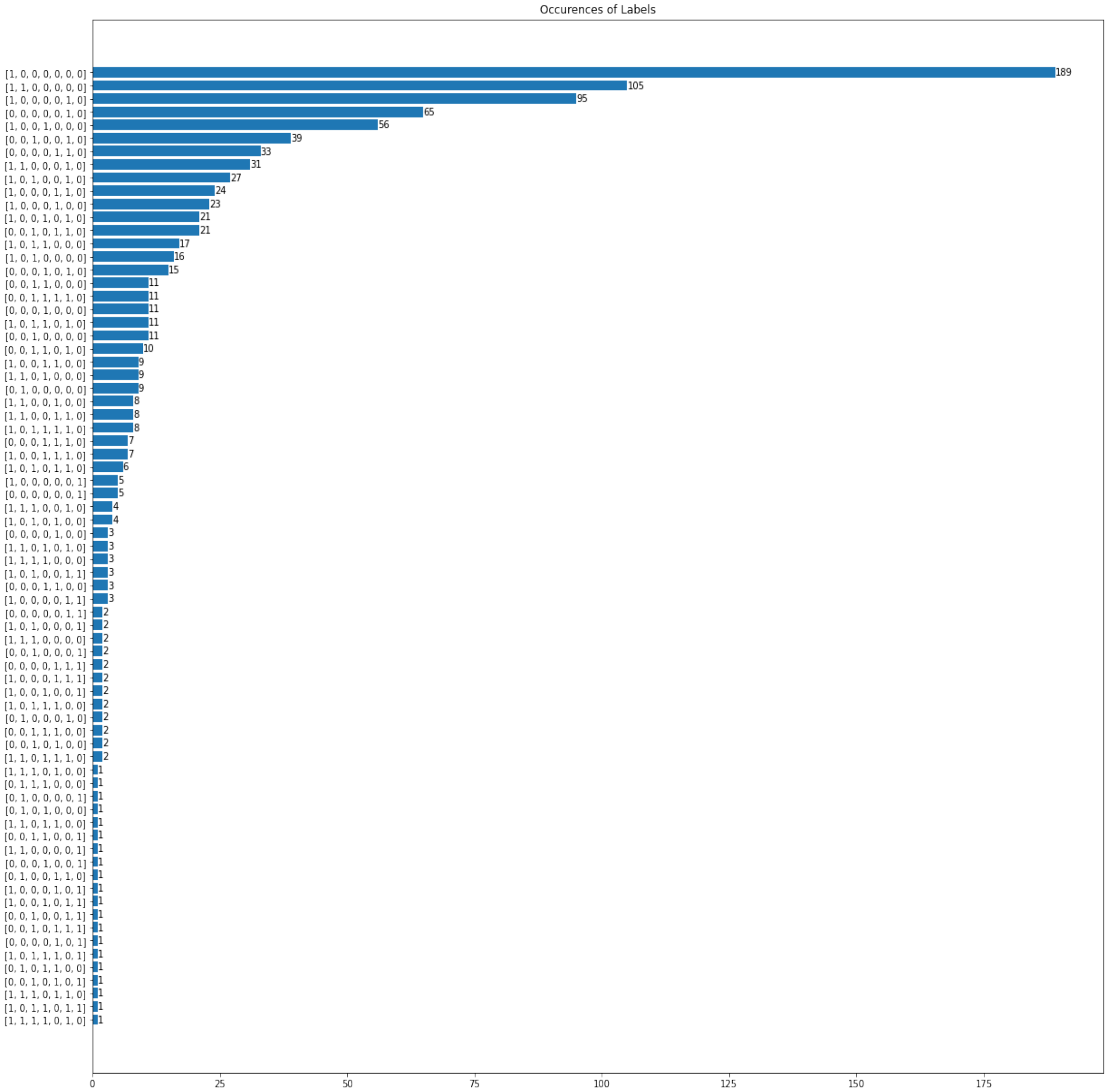

Figure 5: Distribution of label sets in the train data.

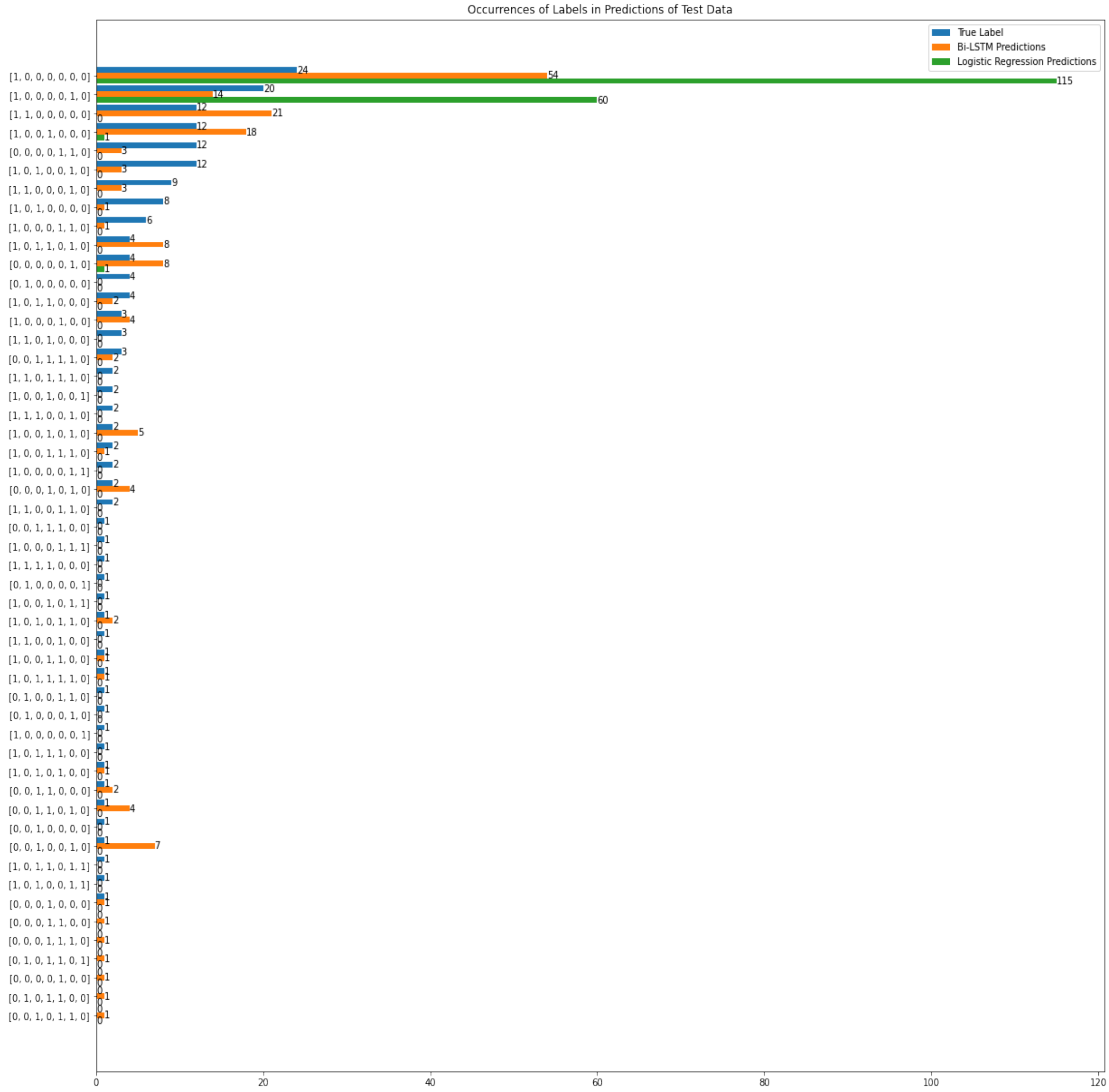

Figure 6: Distribution of label sets in the test data and the predictions on this data by the logistic regression and BiLSTM.

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 12: 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.

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 13: 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.

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

	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 15: 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.