Corpus of Opinion Articles analysis

Piero Castriota

June 19, 2023

Abstract

This project undertakes an in-depth analysis of a database comprising 373 documents, annotated by three distinct annotators. Notably, each article exhibits three different annotations, capturing diverse perspectives. The central objective of this study is to analyse the disparities inherent in these annotations and elucidate their impact on the training and testing of Natural Language Processing (NLP) models. By unraveling the intricacies of these variations, we strive to enhance our comprehension of how divergent perspectives can shape model performance.

1 Introduction

The process of labeling textual data plays a crucial role in numerous natural language processing (NLP) tasks. It involves assigning specific tags or categories to individual units of text, known as Annotated Discourse Units (ADUs), in order to facilitate subsequent analysis and machine learning-based model training. ADUs can range from individual sentences or phrases to entire documents, depending on the task at hand.

One fascinating aspect of the labeling process is the inevitable presence of multiple annotators, each tasked with independently annotating the same set of documents. Despite being provided with the same guidelines and instructions, these annotators often produce annotations that differ from one another. The reasons behind these differences can be attributed to various factors such as varying levels of subjectivity, differences in interpretation, individual biases, or even diverse linguistic backgrounds. Such discrepancies in annotations can arise in both the labeling of categorical information (e.g., sentiment analysis) and more nuanced tasks (e.g., entity recognition or semantic role labeling).

Understanding the differences that arise from multiple annotations is of utmost importance for several reasons. Firstly, it provides insights into the inherent challenges faced by NLP models when dealing with diverse interpretations of text. By analyzing these variations, researchers can gain a deeper understanding of the limitations and pitfalls of relying solely on individual annotations as ground truth. Secondly, understanding the differences between different annotations enables the identification of potential biases or inconsistencies within the labeling process itself. By recognizing and addressing these discrepancies, we can improve the quality and reliability of labeled datasets, leading to more accurate and robust NLP models.

2 Dataset description

- Table 1: Annotation Summary by Annotator
  - Reveals the number of Argumentative Discourse Units (ADUs) and links annotated by each annotator.
- Table 2: Label Summary by Annotator
  - Provides a breakdown of the label distribution for each annotator, indicating the frequency of different argumentative labels assigned.
- Table 3: ADU Lengths by Annotator
  - Displays the average length and maximum length of ADUs annotated by each annotator.

- Table 4: Document Lengths by Annotator
  - Shows the average length and maximum length of documents annotated by each annotator.
- Table 5: ADU and Document Lengths by Annotator
  - Presents the average length and maximum length of both ADUs and documents annotated by each annotator for comparison.
- Table 6: Label Counts by Annotator A: Links
  - Displays the distribution of argumentative labels within the annotations of Annotator A for the "Links" category.
- Table 7: Label Counts by Annotator B: Links
  - Shows the distribution of argumentative labels within the annotations of Annotator B for the "Links" category.
- Table 8: Label Counts by Annotator C: Links
  - Illustrates the distribution of argumentative labels within the annotations of Annotator C for the "Links" category.
- Table 9: Label Counts by Annotator D: Links
  - Highlights the distribution of argumentative labels within the annotations of Annotator D for the "Links" category.

Table 1: Annotation Summary by Annotator

Annotator	**ADUs**	**Links**
A	5444	4273
B	8289	6580
C	6683	5314
D	6593	5247

Table 2: Label Summary by Annotator

Annotator	**Facto**	**Valor(+)**	**nullADU**	**Diretiva**	**Valor**	**Valor(-)**
A	647	183	2109	56	2059	390
B	920	466	3063	167	2790	883
C	386	481	2572	265	2006	973
D	1710	281	2522	179	1247	654

Table 3: ADU Lengths by Annotator

Annotator	**Average Length**	**Max Length**
A	78.15	514
B	90.86	478
C	89.86	546
D	91.72	478

Table 4: Document Lengths by Annotator

Annotator	**Average Length**	**Max Length**
A	2357.39	16428
B	2484.84	18414
C	2993.38	18156
D	2227.13	12947

Table 5: ADU and Document Lengths by Annotator

Annotator	**ADU**		**Documents**	
	Avg. Length	**Max Length**	**Avg. Length**	**Max Length**
A	78.15	514	2357.39	16428
B	90.86	478	2484.84	18414
C	89.86	546	2993.38	18156
D	91.72	478	2227.13	12947

Table 6: Label Counts by Annotator A: Links

Label	**Facto**	**Valor(+)**	**nullADU**	**Diretiva**	**Valor**	**Valor(-)**
Facto	-	-	554	-	-	-
Valor(+)	-	-	103	-	-	-
nullADU	195	155	-	66	1412	281
Diretiva	-	-	7	-	-	-
Valor	-	-	1258	-	-	-
Valor(-)	-	-	242	-	-	-

Table 7: Label Counts by Annotator B: Links

Label	**Facto**	**Valor(+)**	**nullADU**	**Diretiva**	**Valor**	**Valor(-)**
Facto	-	-	800	-	-	-
Valor(+)	-	-	279	-	-	-
nullADU	285	331	-	168	1660	619
Diretiva	-	-	40	-	-	-
Valor	-	-	1830	-	-	-
Valor(-)	-	-	568	-	-	-

Table 8: Label Counts by Annotator C: Links

Label	**Facto**	**Valor(+)**	**nullADU**	**Diretiva**	**Valor**	**Valor(-)**
Facto	-	-	340	-	-	-
Valor(+)	-	-	309	-	-	-
nullADU	104	334	-	287	1154	693
Diretiva	-	-	60	-	-	-
Valor	-	-	1388	-	-	-
Valor(-)	-	-	645	-	-	-

Table 9: Label Counts by Annotator D: Links

Label	**Facto**	**Valor(+)**	**nullADU**	**Diretiva**	**Valor**	**Valor(-)**
Facto	-	-	1732	-	-	-
Valor(+)	-	-	150	-	-	-
nullADU	401	232	-	165	1165	559
Diretiva	-	-	63	-	-	-
Valor	-	-	414	-	-	-
Valor(-)	-	-	366	-	-	-

3 Model Architecture and Training

The training data for this project was obtained by automatically downloading the dataset from the provided link. However, it was observed that the text within the ADUs (Argumentative Discourse Units) was not consistently well delimited. In some cases, even words were cut in half due to unclear boundaries. Despite numerous attempts to identify a potential shift or pattern in the annotation, no consistent pattern or systematic shift was found. This posed a challenge in accurately capturing and labeling the ADUs, requiring additional attention and careful handling during the training process.

The chosen model for this study was DistilBERT Case Uncased, which offers a balance between compactness and speed compared to the larger BERT model while still delivering satisfactory performance. The model underwent training and testing on various datasets to evaluate its effectiveness. During training, the best epoch was retained for each run. However, due to resource constraints, the model was trained for a limited number of epochs, specifically five epochs for each training iteration. The training configurations considered encompassed different combinations of annotators:

- Training on annotator B's data
- Training on annotator D's data
- Training on a combination of annotators A and D's data
- Training on data from all four annotators

Table 10: Training and Validation Results

Training Configuration	**Training Loss**	**Validation Loss**	**Validation Accuracy**
Only D	1.3385	1.3398	0.4506
Only B	1.1329	1.2431	0.5435
Only A and D	1.2434	1.2569	0.4692
All Annotators	1.1483	1.2792	0.4848

4 Results

Table 11: Test Losses

Model	**D**	**B**	**AD**	**All**
Only D	1.3037	1.4389	-	-
Only B	-	1.2264	-	-
Only D and B	-	-	-	-
Only AD	1.2764	1.2708	1.2577	-
All Annotators	1.1954	-	-	1.2853

Table 12: Test Accuracies

Model	**D**	**B**	**AD**	**All**
Only D	0.4595	0.3619	-	-
Only B	-	0.5498	-	-
Only D and B	-	-	-	-
Only AD	0.4676	0.5233	0.5076	-
All Annotators	0.4854	-	-	0.4899

5 Error analysis

A possible reason for the suboptimal results achieved could be attributed to the lack of well-delimited data. With more accurately delimited data, it is anticipated that the model's performance would improve significantly. Clear and precise boundaries for the text within the ADUs are crucial for accurate annotation and subsequent analysis. By ensuring better data delimitation, the model can better capture the context and structure of the arguments, leading to more robust and reliable results. Therefore, efforts should be made to enhance the data delimitation process in order to achieve better overall performance.

6 Conclusions

In general, despite the low accuracies resulting from limited computational resources and poor data quality, some patterns can still be observed. One of the notable findings is the drop in accuracy when a model trained on labels by annotator D is tested on test data created by annotator D and subsequently on data created by annotator B. Conversely, a significant improvement in accuracy is observed when a model trained on labels by annotator B is tested on test data created by annotator B. These observations suggest the presence of annotator-specific patterns and biases that influence the model's performance. The performance variation highlights the importance of considering the annotator's influence and the need for consistent and reliable annotations to ensure robust model training and evaluation.

Additionally, it is worth noting that the accuracy on test data created by annotator D improves as more training data is added. Specifically, the accuracy increases when moving from training solely on data labeled by annotator D to training on data labeled by annotators A and D combined, and further to training on data labeled by all four annotators. This observation suggests that incorporating diverse perspectives and annotations from multiple annotators can lead to improved model performance on test data. It highlights the benefits of leveraging a larger and more diverse training dataset to capture a wider range of patterns and nuances present in the data, ultimately enhancing the model's ability to generalize and make accurate predictions

On the other hand, an interesting observation is that the model trained solely on data labeled by annotators A and D performs better on the test set created by annotator B. This result raises questions about the underlying factors influencing the model's performance. It suggests that there might be certain patterns or similarities between the annotations of annotators A and D that align well with the test data labeled by annotator B. Further investigation is required to understand the specific characteristics of the training data and the interplay between different annotators that contribute to this improved performance. It could be attributed to shared labeling criteria, similar annotation styles, or overlapping perspectives captured by annotators A and D.

7 References

1. DARGMINTS op-articles-arg-pt GitHub repository. Available at: https://github.com/DARGMINTS/op-articles-arg-pt
2. DARGMINTS op-articles-arg-pt Corpus JSON file. Available at: https://raw.githubusercontent.com/DARGMINTS/op-articles-arg-pt/main/op_articles_arg_pt_corpus.json
3. DARGMINTS paper. Available at: //www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.201.pdf