3/9/2021

Machine Learning and Content Analytics

MINI-PROJECT

VASILOPOULOU-CHONDROU ELESA P2822028
VLACHANTONI EVANGELIA P2922002
GKOTSI MARIA-ELENI P2822029

Table of contents

1.Introduction……………………………………………………………………………..……..2
1.1. A few words for “Machine Learning team project”………………..…………….....2
1.2 What is the main idea………………………………………………………..……......2
1.3. What are our goals and expectations from our research?....................................2
2. Methodology………………………………………………………………………………2
2.1. Data Collection…………………………………………………………...….......2
2.2. Description of the data………………………………………………...…........3
2.3 Data processing…………………………………………..………..…..........5
2.4 Annotation…………………………………………………………........……6
3. Experiments in Argument/Structure Prediction……………..…………….………………......6
3.1 Intuitive Baseline ………..………………………….………….……...…………6
3.2 Feed Forward Network Method ……………..………………………….……........8
3.2.1. Argument Prediction…….……………………………….……………...........8
3.2.2. Structure Prediction………………………… ………………........13
3.3 Abstract Clustering……………………………………………………………..16
3.3.1. K-Means…………………………………………………………...……16
3.3.1.a Elbow method………………………………………………..16
3.3.2.b Silhouette method……………………………………….....18
4. Results and Quantitave Analysis……………………………………………….……….....22
4.1 Visualizations……………………………..…..…………………………….....22
4.2 Quantitave Analysis and Errors………………………………...…………....22
5. Discussion and Future goals…………………………………………..……………..23
6. Members and Roles……………………………………………………………….........23
7. Time Plan……………………………………………………………………………......23
8. Table of figures…………………………………………………………………...24
9. Bibliography………………………………………………………………………...25

1. Introduction

1.1. A few words for our "Machine Learning team project"

Our team was established in July 2021 for the needs of Machine Learning mini-project and consists of three members, all of them women, each of one aiming to contribute from her point of view and field of study in this accomplishment. More specifically, Mrs Vassilopoulou-Chondrou Elessa, as Statistician and Business Analyst Consultant, Mrs Vlachantoni Evangelia, Economist and Payroll manager and Mrs Maria-Eleni Gkotsi, Software Engineer and CRM Consultant.

1.2. What is the main idea

The main idea of this project is to predict structure or argument labels when examining abstracts of published articles of different scientific fields. Project's approach will focus first on creating a baseline intuition that will classify each line of abstract to a specific argument label and then by the use of neural networks will try to create a model that predicts argument and structure labels. There are many ways to proceed into this venture. Three main ways are Fasttext Approach (Version 1 and 2), Custom Approach and finally Transformer Network Approach one of which will be applied and analyzed in depth. The final step will be Abstract Clustering with the use of k-means method, trying to reveal main groups of scientific abstract's content.

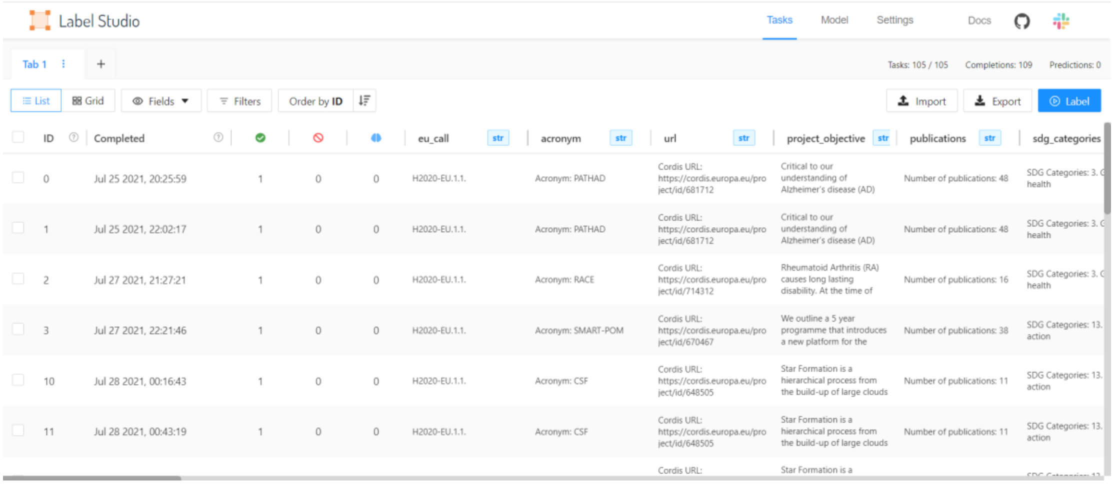
1.3. What are our goals and expectations from our research

Through this research we expect to discover the power of neural networks more specific the application on text recognition, to enrich our skills in Python Programming language and to learn from each other. We expect to create a model that predicts structure and argument labels accurately with low level of error, to depict the groups of scientific abstracts and finally to compare our results with our baseline intuition created as first step.

2. Methodology

2.1 Data Collection

All abstracts were collected and uploaded in Label Studio (Figure 1) where we conducted annotation procedure in over 100 scientific abstracts without collaboration only by following tutor's instructions about main abstract's structure and basic rules for this kind of text annotation. This phase of methodology, was a necessary step in order to create our dataset in following steps.



Tasks: 105 / 105 Completions: 109 Predictions: 0

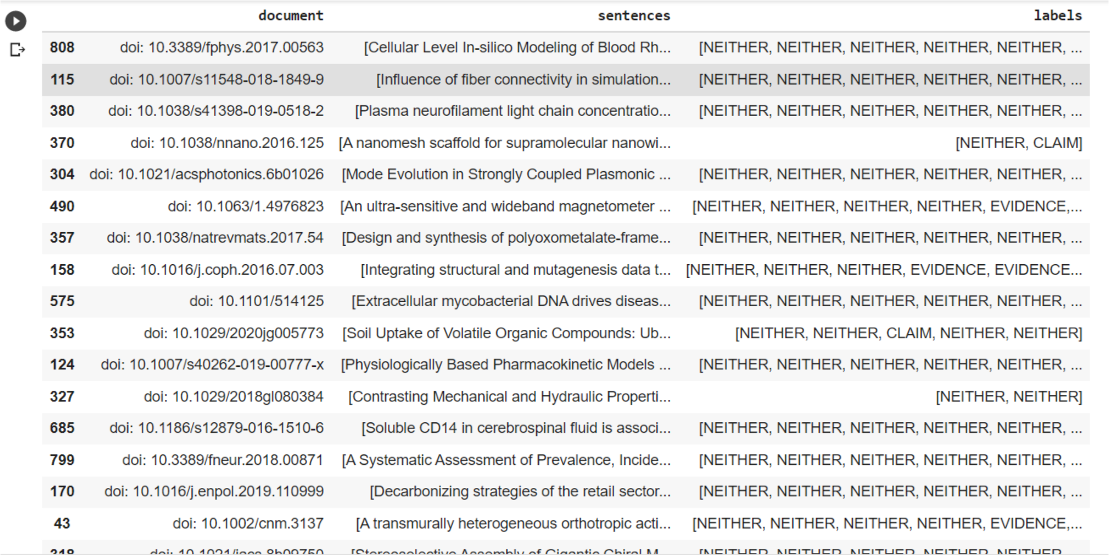
□	ID	Completed	✓	⊘	●	eu_call	acronym	url	project_objective	publications	sdg_categories
□	0	Jul 25 2021, 20:25:59	1	0	0	H2020-EU.1.1.	Acronym: PATHAD	Cordis URL: https://cordis.europa.eu/project/id/681712	Critical to our understanding of Alzheimer's disease (AD)	Number of publications: 48	SDG Categories: 3. G health
□	1	Jul 25 2021, 22:02:17	1	0	0	H2020-EU.1.1.	Acronym: PATHAD	Cordis URL: https://cordis.europa.eu/project/id/681712	Critical to our understanding of Alzheimer's disease (AD)	Number of publications: 48	SDG Categories: 3. G health
□	2	Jul 27 2021, 21:27:21	1	0	0	H2020-EU.1.1.	Acronym: RACE	Cordis URL: https://cordis.europa.eu/project/id/714312	Rheumatoid Arthritis (RA) causes long lasting disability. At the time of	Number of publications: 16	SDG Categories: 3. G health
□	3	Jul 27 2021, 22:21:46	1	0	0	H2020-EU.1.1.	Acronym: SMART-POM	Cordis URL: https://cordis.europa.eu/project/id/670467	We outline a 5 year programme that introduces a new platform for the	Number of publications: 38	SDG Categories: 13. action
□	10	Jul 28 2021, 00:16:43	1	0	0	H2020-EU.1.1.	Acronym: CSF	Cordis URL: https://cordis.europa.eu/project/id/648505	Star Formation is a hierarchical process from the build-up of large clouds	Number of publications: 11	SDG Categories: 13. action
□	11	Jul 28 2021, 00:43:19	1	0	0	H2020-EU.1.1.	Acronym: CSF	Cordis URL: https://cordis.europa.eu/project/id/648505	Star Formation is a hierarchical process from the build-up of large clouds	Number of publications: 11	SDG Categories: 13. action
								Cordis URL:	Star Formation is a		

Figure 1 Data Collection in the Label Studio

2.2 Description of the data

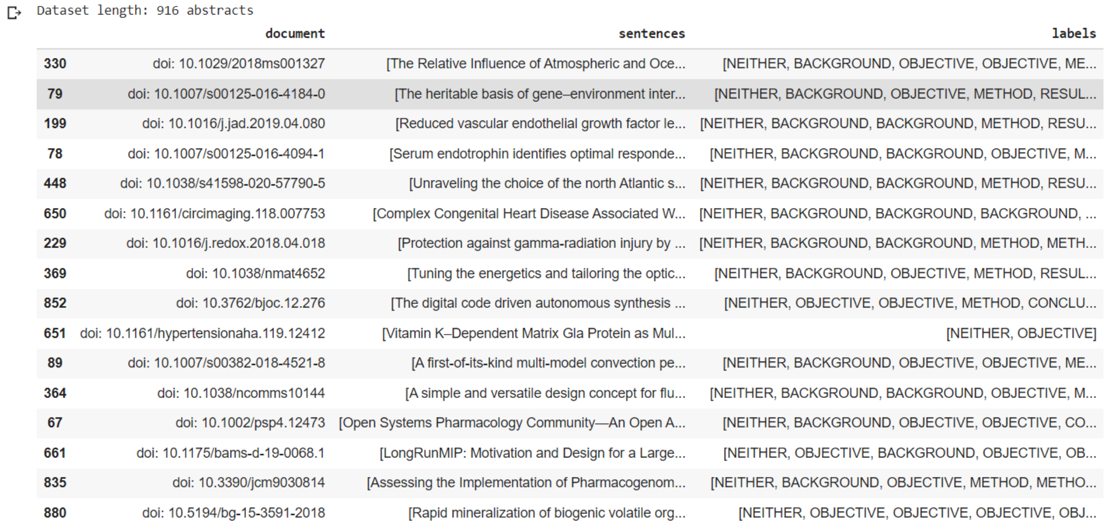
Our data splits into three different datasets. The first one is containing argument labels (Figure 2), the second one structure labels (Figure 3) and the last one citations labels (Figure 4). All three datasets contain three columns named Document, Sentences and Labels. The first column of each dataset contains the doi number of each abstract, the second column contains the sentences and last column the labels of each category. For example, the labels of argument are divided into two subcategories: a) Claim, b) Evidence. Structure's labels are divided into five different subcategories: a) Background, b) Objective/Aim c) Method d) Result e) Conclusion. Depending on the keywords we choose to annotate each sentence with a specific structure label. The basic structure of a scientific abstract is the one mentioned above from a to e. As far as concerned argument labels

were combined either with conclusion or result label (claim and evidence respectively). If there were not any claim in the abstract it was difficult to find evidence.



	document	sentences	labels
808	doi: 10.3389/fphys.2017.00563	[Cellular Level In-silico Modeling of Blood Rh...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
115	doi: 10.1007/s11548-018-1849-9	[Influence of fiber connectivity in simulation...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
380	doi: 10.1038/s41398-019-0518-2	[Plasma neurofilament light chain concentratio...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
370	doi: 10.1038/nnano.2016.125	[A nanomesh scaffold for supramolecular nanowi...	[NEITHER, CLAIM]
304	doi: 10.1021/acsphotonics.6b01026	[Mode Evolution in Strongly Coupled Plasmonic ...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
490	doi: 10.1063/1.4976823	[An ultra-sensitive and wideband magnetometer ...	[NEITHER, NEITHER, NEITHER, NEITHER, EVIDENCE,...
357	doi: 10.1038/natrevmats.2017.54	[Design and synthesis of polyoxometalate-frame...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
158	doi: 10.1016/j.coph.2016.07.003	[Integrating structural and mutagenesis data t...	[NEITHER, NEITHER, NEITHER, EVIDENCE, EVIDENCE...
575	doi: 10.1101/514125	[Extracellular mycobacterial DNA drives diseas...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
353	doi: 10.1029/2020jg005773	[Soil Uptake of Volatile Organic Compounds: Ub...	[NEITHER, NEITHER, CLAIM, NEITHER, NEITHER]
124	doi: 10.1007/s40262-019-00777-x	[Physiologically Based Pharmacokinetic Models ...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
327	doi: 10.1029/2018gl080384	[Contrasting Mechanical and Hydraulic Properti...	[NEITHER, NEITHER]
685	doi: 10.1186/s12879-016-1510-6	[Soluble CD14 in cerebrospinal fluid is associ...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
799	doi: 10.3389/fneur.2018.00871	[A Systematic Assessment of Prevalence, Incide...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
170	doi: 10.1016/j.enpol.2019.110999	[Decarbonizing strategies of the retail sector...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, ...
43	doi: 10.1002/cnm.3137	[A transmurally heterogeneous orthotropic acti...	[NEITHER, NEITHER, NEITHER, NEITHER, EVIDENCE,...
318	doi: 10.1021/jacs.8b09750	[Stereoselective Assembly of Gigantic Chiral M...	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER

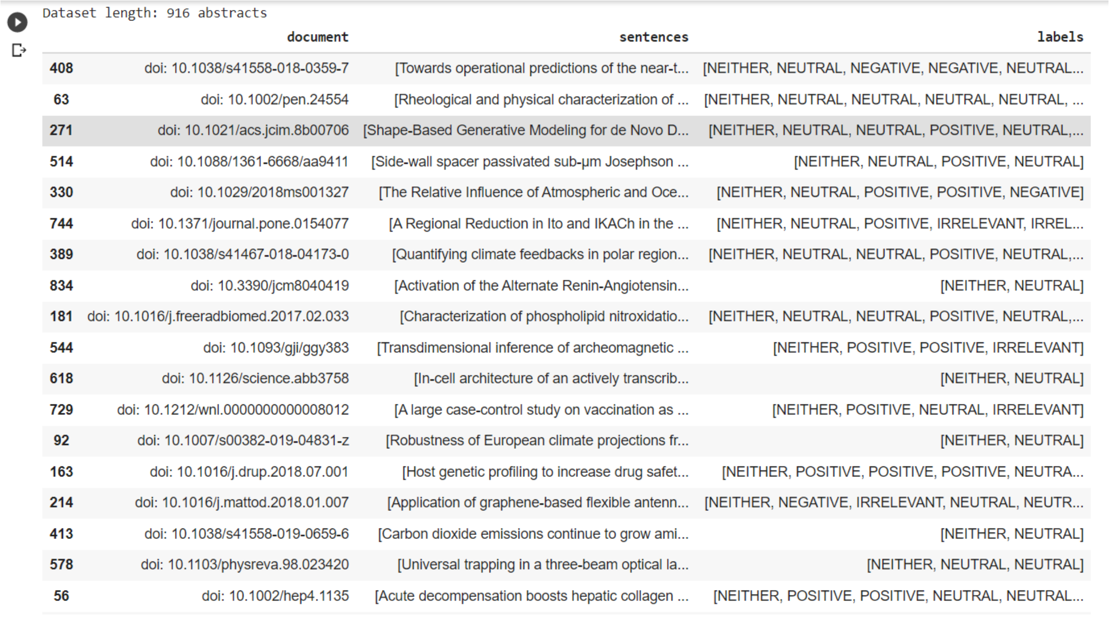
Figure 2 Argument dataset



```
Dataset length: 916 abstracts
```

	document	sentences	labels
330	doi: 10.1029/2018ms001327	[The Relative Influence of Atmospheric and Oce...	[NEITHER, BACKGROUND, OBJECTIVE, OBJECTIVE, ME...
79	doi: 10.1007/s00125-016-4184-0	[The heritable basis of gene–environment inter...	[NEITHER, BACKGROUND, OBJECTIVE, METHOD, RESUL...
199	doi: 10.1016/j.jad.2019.04.080	[Reduced vascular endothelial growth factor le...	[NEITHER, BACKGROUND, BACKGROUND, METHOD, RESU...
78	doi: 10.1007/s00125-016-4094-1	[Serum endotrophin identifies optimal responde...	[NEITHER, BACKGROUND, BACKGROUND, OBJECTIVE, M...
448	doi: 10.1038/s41598-020-57790-5	[Unraveling the choice of the north Atlantic s...	[NEITHER, BACKGROUND, BACKGROUND, METHOD, RESU...
650	doi: 10.1161/circimaging.118.007753	[Complex Congenital Heart Disease Associated W...	[NEITHER, BACKGROUND, BACKGROUND, BACKGROUND, ...
229	doi: 10.1016/j.redox.2018.04.018	[Protection against gamma-radiation injury by ...	[NEITHER, BACKGROUND, BACKGROUND, METHOD, METH...
369	doi: 10.1038/nmat4652	[Tuning the energetics and tailoring the optic...	[NEITHER, BACKGROUND, OBJECTIVE, METHOD, RESUL...
852	doi: 10.3762/bjoc.12.276	[The digital code driven autonomous synthesis ...	[NEITHER, OBJECTIVE, OBJECTIVE, METHOD, CONCLU...
651	doi: 10.1161/hypertensionaha.119.12412	[Vitamin K–Dependent Matrix Gla Protein as Mul...	[NEITHER, OBJECTIVE]
89	doi: 10.1007/s00382-018-4521-8	[A first-of-its-kind multi-model convection pe...	[NEITHER, BACKGROUND, OBJECTIVE, OBJECTIVE, ME...
364	doi: 10.1038/ncomms10144	[A simple and versatile design concept for flu...	[NEITHER, BACKGROUND, BACKGROUND, OBJECTIVE, M...
67	doi: 10.1002/psp4.12473	[Open Systems Pharmacology Community—An Open A...	[NEITHER, BACKGROUND, OBJECTIVE, OBJECTIVE, CO...
661	doi: 10.1175/bams-d-19-0068.1	[LongRunMIP: Motivation and Design for a Large...	[NEITHER, OBJECTIVE, BACKGROUND, OBJECTIVE, OB...
835	doi: 10.3390/jcm9030814	[Assessing the Implementation of Pharmacogenom...	[NEITHER, BACKGROUND, OBJECTIVE, METHOD, METHO...
880	doi: 10.5194/bg-15-3591-2018	[Rapid mineralization of biogenic volatile org...	[NEITHER, OBJECTIVE, OBJECTIVE, OBJECTIVE, OBJ...

Figure 3 Structure dataset



Dataset length: 916 abstracts

	document	sentences	labels
408	doi: 10.1038/s41558-018-0359-7	[Towards operational predictions of the near-t...	[NEITHER, NEUTRAL, NEGATIVE, NEGATIVE, NEUTRAL...
63	doi: 10.1002/pen.24554	[Rheological and physical characterization of ...	[NEITHER, NEUTRAL, NEUTRAL, NEUTRAL, NEUTRAL, ...
271	doi: 10.1021/acs.jcim.8b00706	[Shape-Based Generative Modeling for de Novo D...	[NEITHER, NEUTRAL, NEUTRAL, POSITIVE, NEUTRAL,...
514	doi: 10.1088/1361-6668/aa9411	[Side-wall spacer passivated sub-μm Josephson ...	[NEITHER, NEUTRAL, POSITIVE, NEUTRAL]
330	doi: 10.1029/2018ms001327	[The Relative Influence of Atmospheric and Oce...	[NEITHER, NEUTRAL, POSITIVE, POSITIVE, NEGATIVE]
744	doi: 10.1371/journal.pone.0154077	[A Regional Reduction in Ito and IKACh in the ...	[NEITHER, NEUTRAL, POSITIVE, IRRELEVANT, IRREL...
389	doi: 10.1038/s41467-018-04173-0	[Quantifying climate feedbacks in polar region...	[NEITHER, NEUTRAL, NEUTRAL, POSITIVE, NEUTRAL,...
834	doi: 10.3390/jcm8040419	[Activation of the Alternate Renin-Angiotensin...	[NEITHER, NEUTRAL]
181	doi: 10.1016/j.freeradbiomed.2017.02.033	[Characterization of phospholipid nitroxidatio...	[NEITHER, NEUTRAL, NEUTRAL, POSITIVE, NEUTRAL,...
544	doi: 10.1093/gji/ggy383	[Transdimensional inference of archeomagnetic ...	[NEITHER, POSITIVE, POSITIVE, IRRELEVANT]
618	doi: 10.1126/science.abb3758	[In-cell architecture of an actively transcrib...	[NEITHER, NEUTRAL]
729	doi: 10.1212/wnl.0000000000008012	[A large case-control study on vaccination as ...	[NEITHER, POSITIVE, NEUTRAL, IRRELEVANT]
92	doi: 10.1007/s00382-019-04831-z	[Robustness of European climate projections fr...	[NEITHER, NEUTRAL]
163	doi: 10.1016/j.drup.2018.07.001	[Host genetic profiling to increase drug safet...	[NEITHER, POSITIVE, POSITIVE, POSITIVE, NEUTRA...
214	doi: 10.1016/j.mattod.2018.01.007	[Application of graphene-based flexible antenn...	[NEITHER, NEGATIVE, IRRELEVANT, NEUTRAL, NEUTR...
413	doi: 10.1038/s41558-019-0659-6	[Carbon dioxide emissions continue to grow ami...	[NEITHER, NEUTRAL]
578	doi: 10.1103/physreva.98.023420	[Universal trapping in a three-beam optical la...	[NEITHER, NEUTRAL, NEUTRAL]
56	doi: 10.1002/hep4.1135	[Acute decompensation boosts hepatic collagen ...	[NEITHER, POSITIVE, POSITIVE, NEUTRAL, NEUTRAL...

Figure 4 Citations dataset

Citation labels are divided in four different categories a) Positive b) Negative c) Neutral d) Irrelevant. The aim in this type of label is to characterize whether a citation supports (positive label) or not (negative label) the opinion or scientific analysis exposed in each abstract. With label irrelevant we characterized either citations that were not aligned with their appropriate citation number or citations written in different language or sentences containing symbols with no structure and similar sense.

2.3 Data processing

First step in our data process is to import json files with the corresponding python code. In order to start answering the first question regarding the baseline tuition, we had first to assign sentence column in separate object to create a dataframe easier to read and ready to use it for the first task. The primitive data type of each row of the column sentence was organized in list. Which means each row of sentence is a list. By using explode we created a data frame containing each sentence of type object with a correspondent document id. The same transformation happened to column labels and at the end a new data frame was prepared, named datargument, ready to be used.

2.4 Annotation

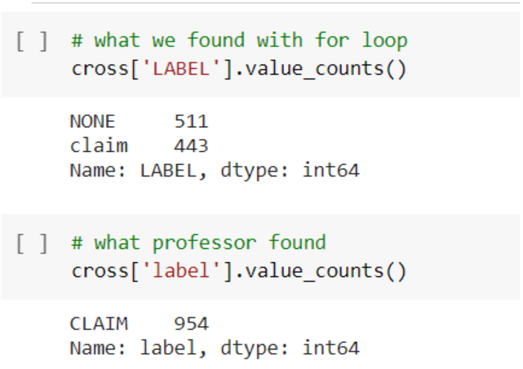
Before uploading our json dataset, in order to accomplish text recognition in upcoming assignment tasks we had to complete a previous important phase. Annotation. Among various scientific abstracts of different subjects, we had to append the appropriate label according three categories. Firstly, the structure label, whether a sentence belongs to background, previous knowledge acquired from other surveys, objective or aim of the scientific abstract, method, results and conclusion. After that, the argument label, whether a sentence already labeled with one of the previous structure could maintain a second label, specifically of evidence or claim depending the content of phrase and keywords. Finally, the last step in the annotation of each abstract, the citation's label, selecting whether a citation was positive, negative, neutral or irrelevant to the writer's analysis.

3. Experiments in Argument/Structure Prediction

3.1 Intuitive Baseline

Digging into our venture we will try to build a simple baseline based on exploratory analysis that will examine the data, for example label distribution in abstracts by using lexicon as a mean of classification that assigns labels to sentences. Our baseline intuition will be applied in argument labels forming an iterative repetition in each sentence which will examine whether there is or not a keyword justifying whether the argument label belongs to claim or evidence.

Before this, an important step to this approach is to create a bug of keywords recently appeared in each argument label category. For this, with function str.split(" ") we create a dataframe with one column containing all abstract's words eliminate the factor of lower case or upper case letters, delete special characters such as commas, points etc. dropping out all duplicates, and create the list that the for loop will use to check if the sentence of the abstract is containing one of them. In the case that one word included in the list exists also in the sentence the corresponding label will be assigned.

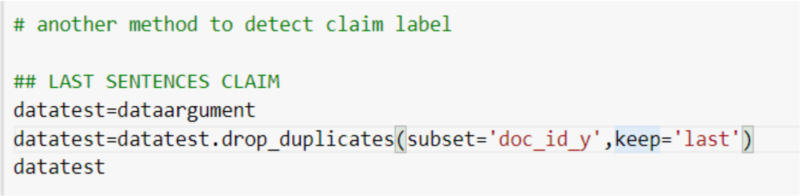


```
[ ]  # what we found with for loop
     cross['LABEL'].value_counts()

     NONE     511
     claim    443
     Name: LABEL, dtype: int64

[ ]  # what professor found
     cross['label'].value_counts()

     CLAIM    954
     Name: label, dtype: int64
```

Figure 5 Cross check Argument Label Claim

After counting what our for loop found as claim label, we saw that from 954 sentences labeled as CLAIM we detected 443 (Figure 5), which means 46,4% of claim labels were assigned correctly, while the rest was not detected by the for loop. The total number of appended label claims using the for loop was 3997 which means 3554 labels were wrongly appended (not very encouraging).

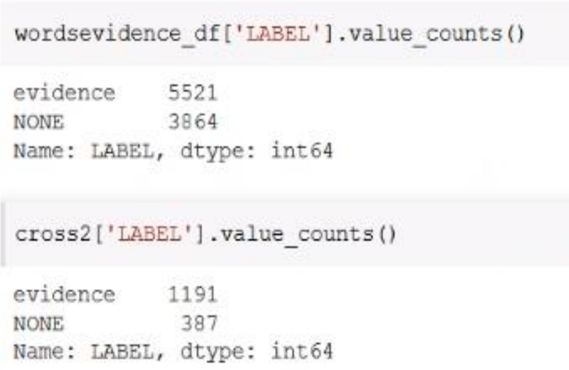
Another approach for the labeling of Claim was to label the last sentences of an Abstract as Claim, because the majority of the last sentences of the Abstracts were Conclusion as Structure and Claim as Argument.



```
# another method to detect claim label

## LAST SENTENCES CLAIM
datatest=dataargument
datatest=datatest.drop_duplicates(subset='doc_id_y',keep='last')
datatest
```

The result here is that we have correctly appended 473 Claim Labels from the total 954 and the rest represent another label.

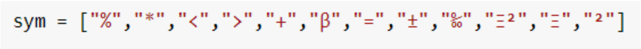
As far as evidence label is concerned, after counting what our for loop found as evidence label, we saw that from 1578 sentences labeled as EVIDENCE we detected 1017 (Figure 6), which means 64,4% of evidence labels where assigned correctly, while the rest was not detected by the for loop. The total of evidence labels obtained from the for loop were 5521 which means that 4330 were wrongly appended (not good as well).



```
wordsevidence_df['LABEL'].value_counts()

evidence    5521
NONE        3864
Name: LABEL, dtype: int64

cross2['LABEL'].value_counts()

evidence    1191
NONE         387
Name: LABEL, dtype: int64
```

Figure 6 Cross check Argument Label Evidence

Another approach for the labeling of the Evidence was to create a list with all the symbols which could be referred to evidence and find the sentences in which these symbols are included.



```
sym = ["%","*","<",">","+","β","=","±","‰","Ξ²","Ξ","²"]
```

The result here was that from the 1578 Evidences, this method found 921 Evidence sentences and from them the 405 were assigned correctly. It seems that this method could be more effective.

3.2 Feed Forward Network Method

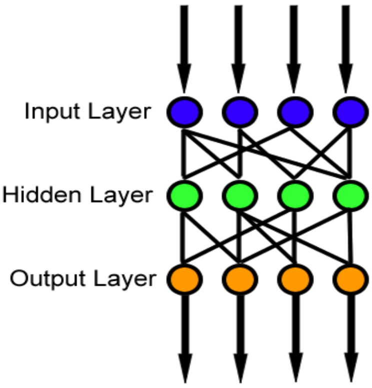
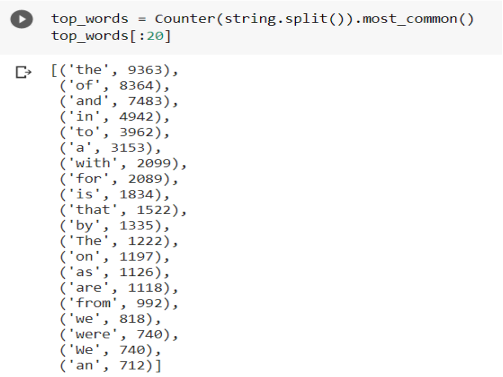


Figure 7 Feed Forward Network Schema

A feedforward neural network is an artificial neural network wherein connections between the nodes do not form a cycle. As such, it is different from its descendant: recurrent neural networks. The feedforward neural network was the first and simplest type of artificial neural network devised. In this network, the information moves in only one direction—forward—from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network.

3.2.1. Argument Prediction

After loading our dataset (dataset_aueb_argument_v1) and proceeding with basic descriptive statistics in order to better understand data's content and we isolated sentences and labels columns into two different variables. Next important step is one of the most frequent steps on a machine learning pipeline is splitting data into training, validation and test sets. The splitting process requires a random shuffle of the data followed by a partition using a preset threshold. On classification variants, you may want to use stratification to ensure the same distribution of classes on both sets. Splitting procedure occurred in two phases. First, we created train-validation and test dataset, and afterwards we continued with splitting train-validation into two separated datasets, the train and the validation one. In order to make the algorithm work, we had to encode our y variable in both three sets by using OneHotEncoder. The idea is to create arrays with 0-1 values representing whether a sentence is characterized as claim, evidence or neither. If one of the three previous options existed as label in a sentence then it was represented in the array as one. In other case 0. The procedure which is known as embeddings is a method used to represent discrete variables as continuous vectors. In the context of neural networks, embeddings are low-dimensional, learned continuous vector representations of discrete variables. Neural network embeddings are useful because they can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space. Next step in our pre-process itinerary is the creation of tokens. As creation of tokens or Tokenization, is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords. Hence, tokenization can be broadly classified into 3 types – word, character, and subword (n-gram characters) tokenization. Creating Vocabulary is the ultimate goal of Tokenization. One of the simplest hacks to boost the performance of the NLP model is to create a vocabulary out of top K frequently occurring words. Now, let's understand the usage of the vocabulary in Traditional and Advanced Deep Learning-based NLP methods. Traditional NLP approaches such as Count Vectorizer and TF-IDF use vocabulary as features. Each word in the vocabulary is treated as a unique feature. In Advanced Deep Learning-based NLP architectures, vocabulary is used to create the tokenized input sentences. Finally, the tokens of these sentences are passed as inputs to the model.

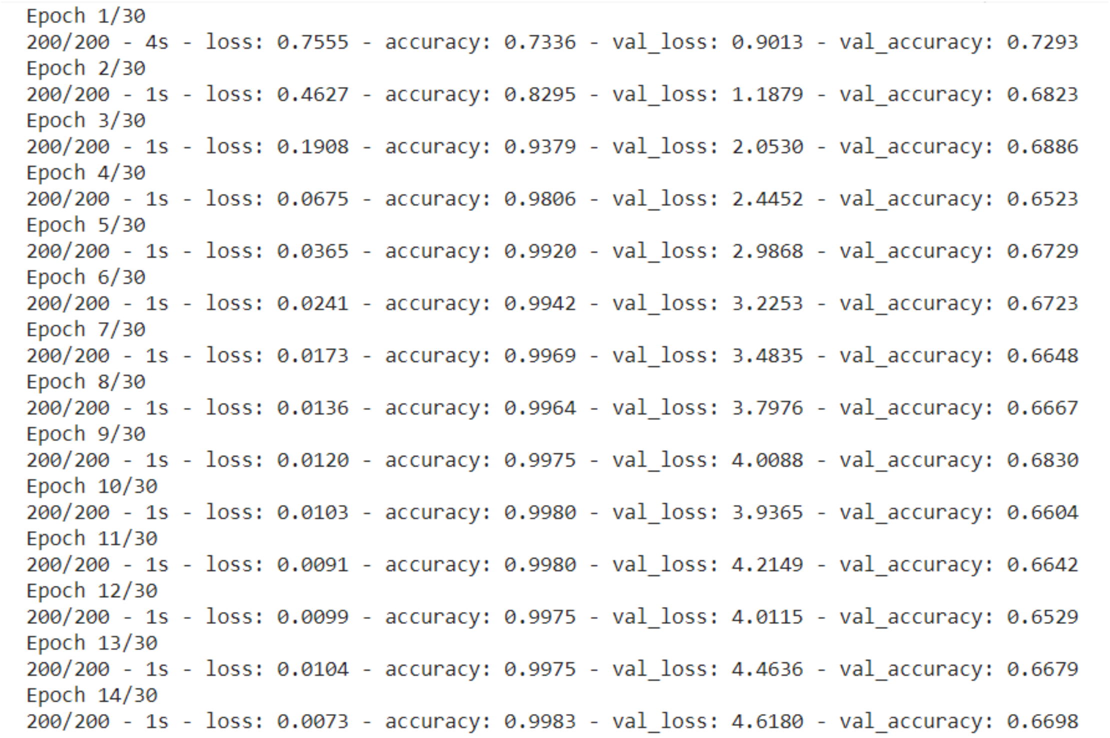


```
top_words = Counter(string.split()).most_common()
top_words[:20]
```

```
[('the', 9363),
 ('of', 8364),
 ('and', 7483),
 ('in', 4942),
 ('to', 3962),
 ('a', 3153),
 ('with', 2099),
 ('for', 2089),
 ('is', 1834),
 ('that', 1522),
 ('by', 1335),
 ('The', 1222),
 ('on', 1197),
 ('as', 1126),
 ('are', 1118),
 ('from', 992),
 ('we', 818),
 ('were', 740),
 ('We', 740),
 ('an', 712)]
```

Figure 8 Count the frequency of top words

Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is underfitted doesn't match closely enough. In our case, regarding the pairs of loss- accuracy for train and validation dataset (Figure 9) we see that while in the train set our accuracy is constantly rising to 0.9983 on the contrary in the validation dataset is starting from 0.7293 keeping a downward trend that stops in 0.6554.



```
Epoch 1/30
200/200 - 4s - loss: 0.7555 - accuracy: 0.7336 - val_loss: 0.9013 - val_accuracy: 0.7293
Epoch 2/30
200/200 - 1s - loss: 0.4627 - accuracy: 0.8295 - val_loss: 1.1879 - val_accuracy: 0.6823
Epoch 3/30
200/200 - 1s - loss: 0.1908 - accuracy: 0.9379 - val_loss: 2.0530 - val_accuracy: 0.6886
Epoch 4/30
200/200 - 1s - loss: 0.0675 - accuracy: 0.9806 - val_loss: 2.4452 - val_accuracy: 0.6523
Epoch 5/30
200/200 - 1s - loss: 0.0365 - accuracy: 0.9920 - val_loss: 2.9868 - val_accuracy: 0.6729
Epoch 6/30
200/200 - 1s - loss: 0.0241 - accuracy: 0.9942 - val_loss: 3.2253 - val_accuracy: 0.6723
Epoch 7/30
200/200 - 1s - loss: 0.0173 - accuracy: 0.9969 - val_loss: 3.4835 - val_accuracy: 0.6648
Epoch 8/30
200/200 - 1s - loss: 0.0136 - accuracy: 0.9964 - val_loss: 3.7976 - val_accuracy: 0.6667
Epoch 9/30
200/200 - 1s - loss: 0.0120 - accuracy: 0.9975 - val_loss: 4.0088 - val_accuracy: 0.6830
Epoch 10/30
200/200 - 1s - loss: 0.0103 - accuracy: 0.9980 - val_loss: 3.9365 - val_accuracy: 0.6604
Epoch 11/30
200/200 - 1s - loss: 0.0091 - accuracy: 0.9980 - val_loss: 4.2149 - val_accuracy: 0.6642
Epoch 12/30
200/200 - 1s - loss: 0.0099 - accuracy: 0.9975 - val_loss: 4.0115 - val_accuracy: 0.6529
Epoch 13/30
200/200 - 1s - loss: 0.0104 - accuracy: 0.9975 - val_loss: 4.4636 - val_accuracy: 0.6679
Epoch 14/30
200/200 - 1s - loss: 0.0073 - accuracy: 0.9983 - val_loss: 4.6180 - val_accuracy: 0.6698
```

Figure 9 Fit argument prediction model over 32 epochs

Loss is a value that represents the summation of errors in our model. It measures how well (or bad) our model is doing. If the errors are high, the loss will be high, which means that the model does not fit well. Otherwise, the lower it is, the better our model works. In our case, for the train set the loss is always below 0 which means that errors are not high and our model works well, while in the validation set is starting 0.9 keeping upward trend that stops in 5.8760 depicting a moderate performance (Figure 10).

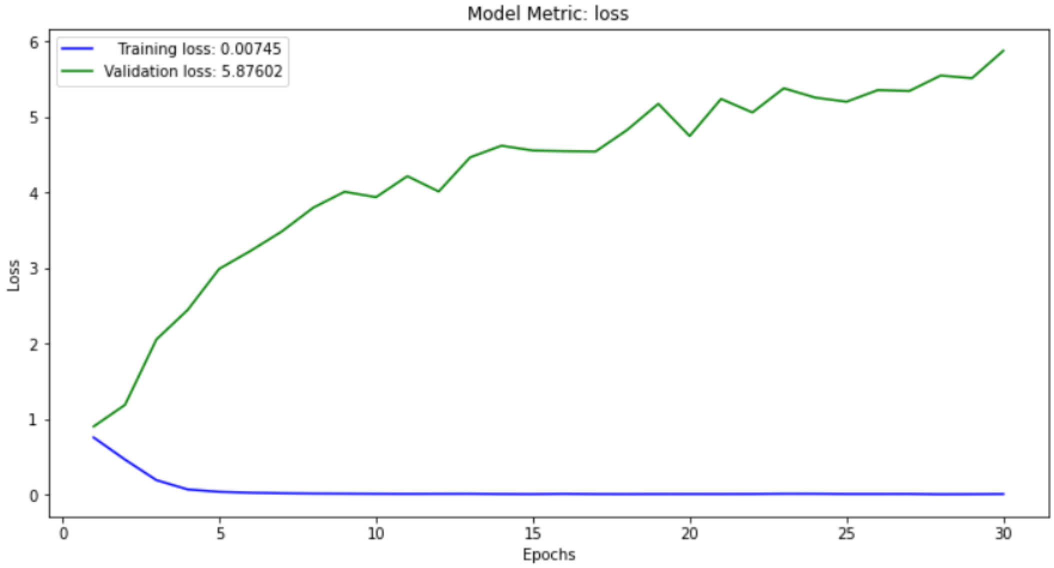


Figure 10 Model metric loss

Accuracy is more straightforward**.** It measures how well our model predicts by comparing the model predictions with the true values in terms of percentage which means that 99,83% our model predicts well on the train set but on the contrary in the validation dataset it predicts with 65,54% success (Figure 11). Now**,** if we analyze these two measurements together, we can infer more information about how is our model working.

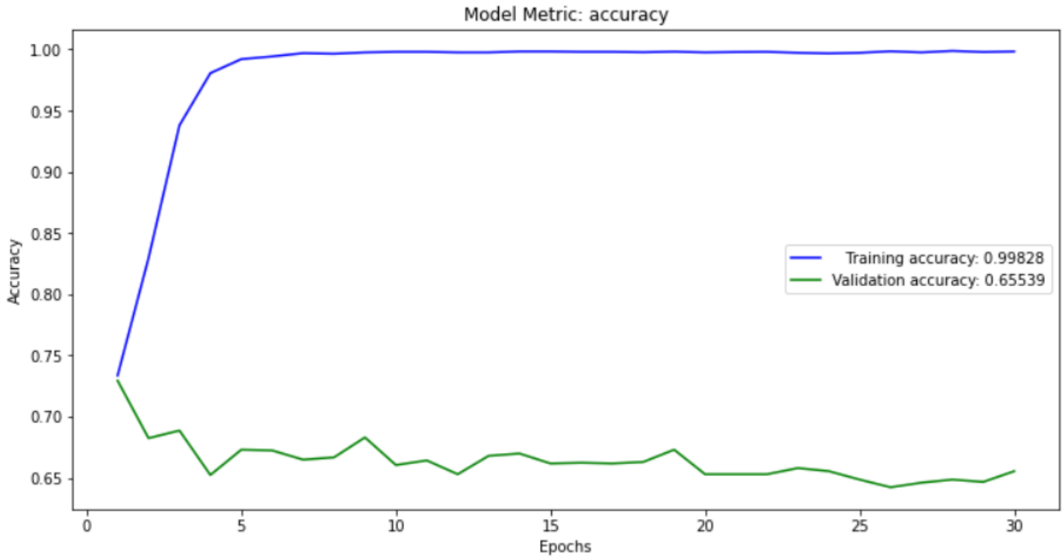
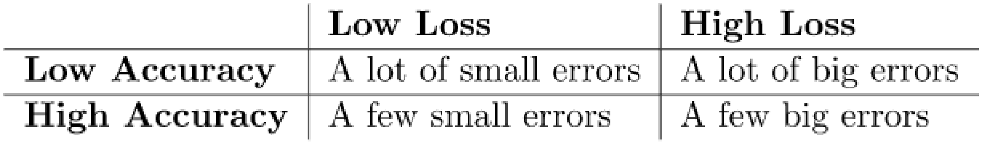


Figure 11 Model metric accuracy

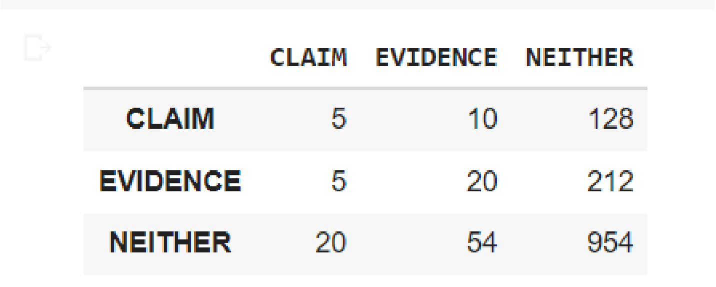
Having a low accuracy but a high loss would mean that the model makes big errors in most of the data. But, if both loss and accuracy are low, it means the model makes small errors in most of the data. However, if they're both high, it makes big errors in some of the data. Finally, if the accuracy is high and the loss is low, then the model makes small errors on just some of the data, which would be the ideal case and is the category that we belong.



	Low Loss	**High Loss**
Low Accuracy	A lot of small errors	A lot of big errors
High Accuracy	A few small errors	A few big errors

Figure 12 Loss-Accuracy interpration

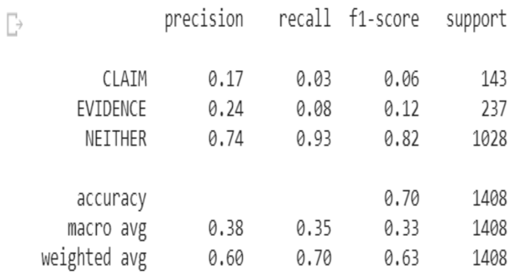
In order to evaluate our model, we are going to implement two methods, the first is confusion matrix and second one classification report. Regarding the confusion matrix, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with different combinations of predicted and actual values.



	CLAIM	EVIDENCE	NEITHER
CLAIM	5	10	128
EVIDENCE	5	20	212
NEITHER	20	54	954

Figure 13 Confusion Matrix Argument Prediction evaluation

On the vertical column we have the predicted and on horizontal we have the actual values. True positive means that we predicted positive and it is true. Here we predict 5 sentences as claim and all five are true. True negative means we predicted negative and it is true. False positive (Type error 1), we predicted positive and it is false. Here we predicted 5 sentences of argument evidence that belonged to argument claim. Finally, false negative (Type 2 Error) we predicted negative and it is false.



```
              precision    recall  f1-score   support

       CLAIM       0.17      0.03      0.06       143
    EVIDENCE       0.24      0.08      0.12       237
     NEITHER       0.74      0.93      0.82      1028

    accuracy                           0.70      1408
   macro avg       0.38      0.35      0.33      1408
weighted avg       0.60      0.70      0.63      1408
```

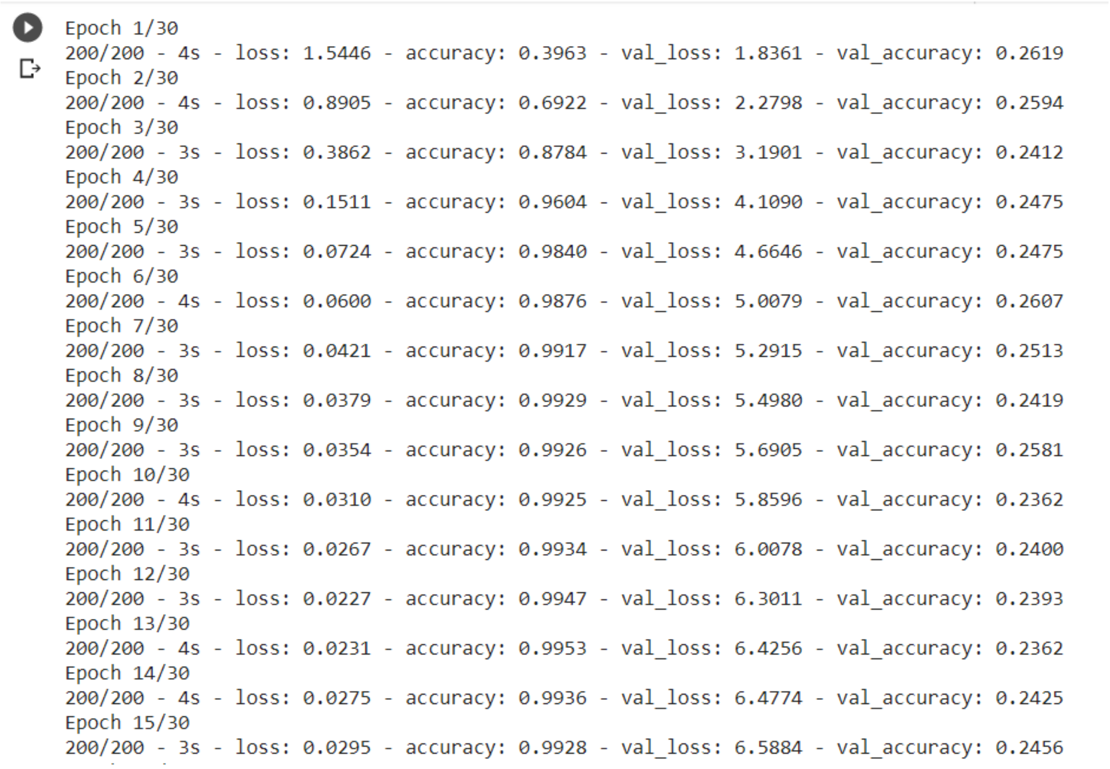
Figure 14 Classification Report Argument Prediction evaluation

The classification report visualizer displays the precision, recall, f1-score, and support scores for the model. Precision — What percent of your predictions were correct

Precision-What percent of your predictions were correct? Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive. Precision: Accuracy of positive predictions. Here the precision for claim is 0.17, for evidence, 0.24 and for neither 0.74 which means that the model predicts better the label “neither”. **Recall -What percent of the positive cases did you catch?** Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. Recall: Fraction of positives that were correctly identified. Here the recall for claim is 0.03, for evidence 0.08 and for neither 0.93 which means that 93% were correct found positive. **F1 score -What percent of positive predictions were correct?** The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. Here f1 for claim is 0.06, for evidence 0.12 and for neither 0.82. Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn’t change between models but instead diagnoses the evaluation process.

3.2.2 Structure Prediction

Further in our analysis, we will examine the model of structure prediction concerning the labels background, objective, method, result conclusion by loading the dataset “dataset_aueb_structure_v1”. The pre-process is the same with the Argument Prediction described in the previous section. After carefully integrate all above steps we tried to fit our model to see the loss-accuracy pair in train and validation dataset.



```
Epoch 1/30
200/200 - 4s - loss: 1.5446 - accuracy: 0.3963 - val_loss: 1.8361 - val_accuracy: 0.2619
Epoch 2/30
200/200 - 4s - loss: 0.8905 - accuracy: 0.6922 - val_loss: 2.2798 - val_accuracy: 0.2594
Epoch 3/30
200/200 - 3s - loss: 0.3862 - accuracy: 0.8784 - val_loss: 3.1901 - val_accuracy: 0.2412
Epoch 4/30
200/200 - 3s - loss: 0.1511 - accuracy: 0.9604 - val_loss: 4.1090 - val_accuracy: 0.2475
Epoch 5/30
200/200 - 3s - loss: 0.0724 - accuracy: 0.9840 - val_loss: 4.6646 - val_accuracy: 0.2475
Epoch 6/30
200/200 - 4s - loss: 0.0600 - accuracy: 0.9876 - val_loss: 5.0079 - val_accuracy: 0.2607
Epoch 7/30
200/200 - 3s - loss: 0.0421 - accuracy: 0.9917 - val_loss: 5.2915 - val_accuracy: 0.2513
Epoch 8/30
200/200 - 3s - loss: 0.0379 - accuracy: 0.9929 - val_loss: 5.4980 - val_accuracy: 0.2419
Epoch 9/30
200/200 - 3s - loss: 0.0354 - accuracy: 0.9926 - val_loss: 5.6905 - val_accuracy: 0.2581
Epoch 10/30
200/200 - 4s - loss: 0.0310 - accuracy: 0.9925 - val_loss: 5.8596 - val_accuracy: 0.2362
Epoch 11/30
200/200 - 3s - loss: 0.0267 - accuracy: 0.9934 - val_loss: 6.0078 - val_accuracy: 0.2400
Epoch 12/30
200/200 - 3s - loss: 0.0227 - accuracy: 0.9947 - val_loss: 6.3011 - val_accuracy: 0.2393
Epoch 13/30
200/200 - 4s - loss: 0.0231 - accuracy: 0.9953 - val_loss: 6.4256 - val_accuracy: 0.2362
Epoch 14/30
200/200 - 4s - loss: 0.0275 - accuracy: 0.9936 - val_loss: 6.4774 - val_accuracy: 0.2425
Epoch 15/30
200/200 - 3s - loss: 0.0295 - accuracy: 0.9928 - val_loss: 6.5884 - val_accuracy: 0.2456
```

Figure 15 Fit structure prediction model over 32 epochs

In comparison to the previous fit of argument prediction here we see that while loss-accuracy pair in train dataset is going well, in the validation loss accuracy pair we do not meet the same results. In other words, in our case above which number the error is considered to be low or big. Here the 6,5 comparing with other kind of errors in statistics we would say that is a small error accompanied with small accuracy, thus in this case our model makes small errors.

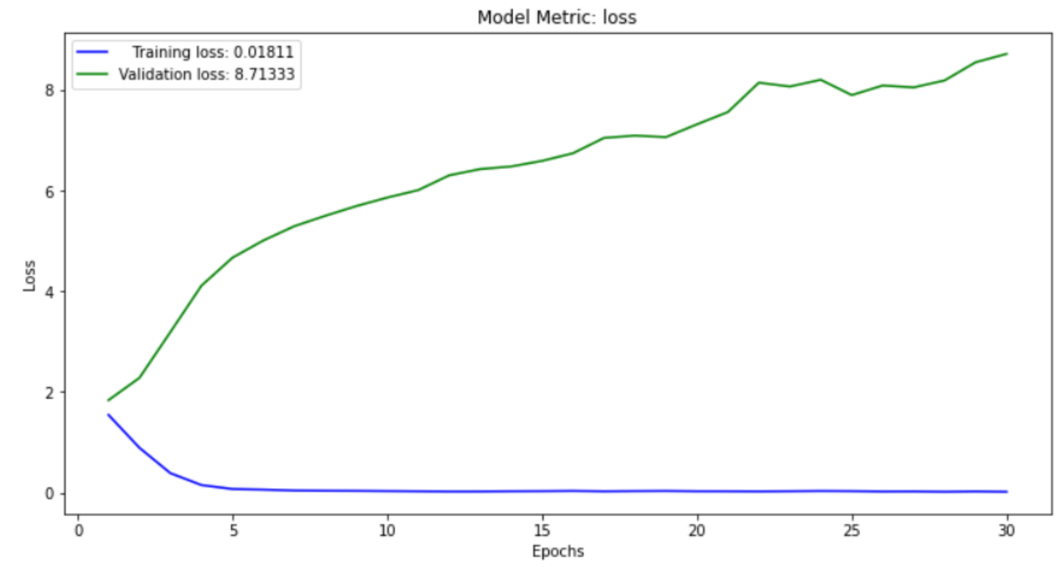


Figure 16 Loss metric structure prediction

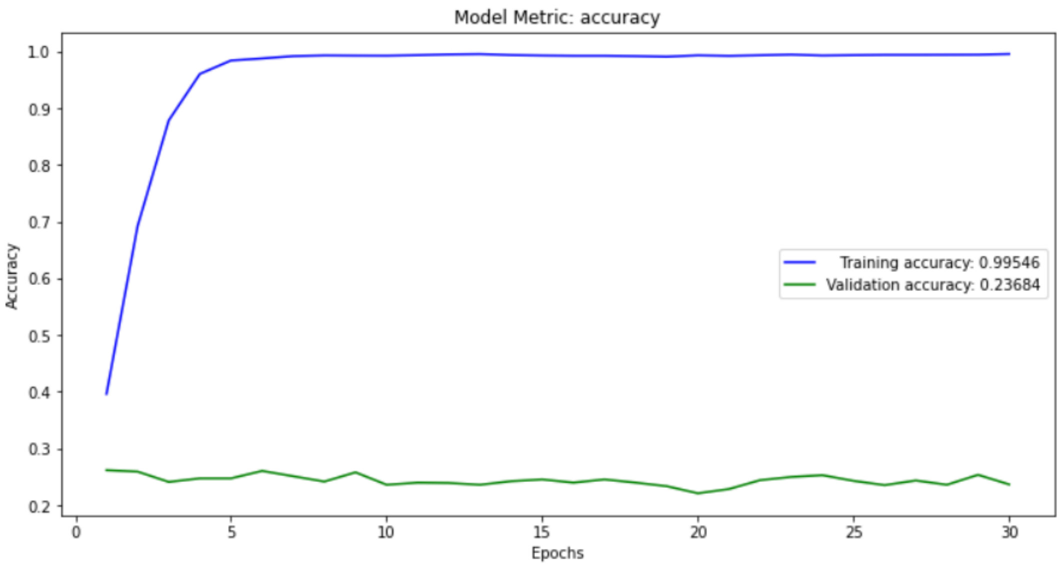
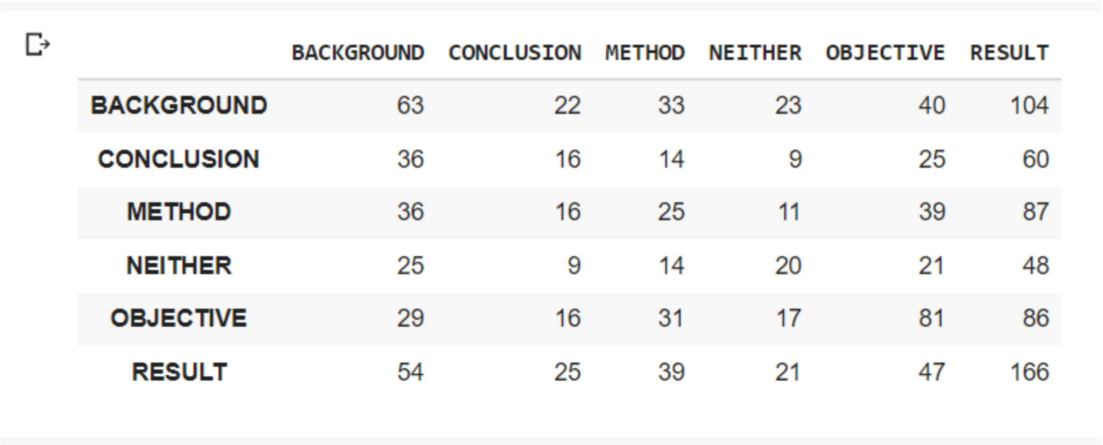
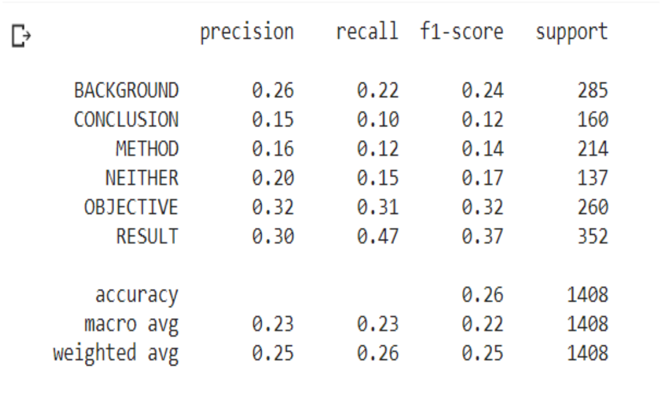


Figure 17 Accuracy structure prediction



	BACKGROUND	CONCLUSION	METHOD	NEITHER	OBJECTIVE	RESULT
BACKGROUND	63	22	33	23	40	104
CONCLUSION	36	16	14	9	25	60
METHOD	36	16	25	11	39	87
NEITHER	25	9	14	20	21	48
OBJECTIVE	29	16	31	17	81	86
RESULT	54	25	39	21	47	166

Figure 18 Confusion Matrix for Structure Prediction



```
              precision    recall  f1-score   support

  BACKGROUND       0.26      0.22      0.24       285
  CONCLUSION       0.15      0.10      0.12       160
      METHOD       0.16      0.12      0.14       214
     NEITHER       0.20      0.15      0.17       137
   OBJECTIVE       0.32      0.31      0.32       260
      RESULT       0.30      0.47      0.37       352

    accuracy                           0.26      1408
   macro avg       0.23      0.23      0.22      1408
weighted avg       0.25      0.26      0.25      1408
```

Figure 19 Classification Report for structure prediction

Structure prediction model provided slightly better results in the prediction of result objective and result label in comparison to other labels but still with many errors. This claim it is identified also by the recall ratio where the 31% for objective is predicted correctly and the 47% for result is predicted correctly as well.

It is important to mention that when we change the epochs from 32 to 60, the performance of the model was improved significantly.

3.3 Abstract Clustering

3.3 K-means

In this part of the assignment, we will use the dataset named “dataset_aueb_argument_v2” which it contains two more columns, the eu_call and the project objective. Our aim is to use any of the embeddings created for Abstract, Project objective (each abstract belongs to a project), EU Call (each project belongs to an EU call), Claim only or Claim and Evidence to create clusters using embeddings from the abstract or a combination of features of any of the above mentioned categories.

3.3.1a Elbow method

The KElbowVisualizer implements the “elbow” method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In our case, we do not meet this point, thus we do not have clusters among abstracts or their combinations with aforementioned features.

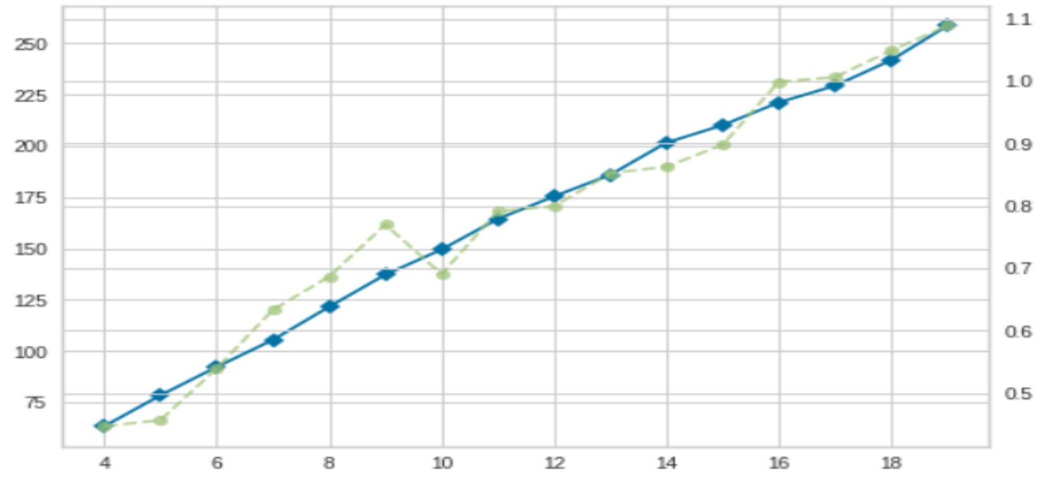


Figure 20 Elbow method for abstract embeddings

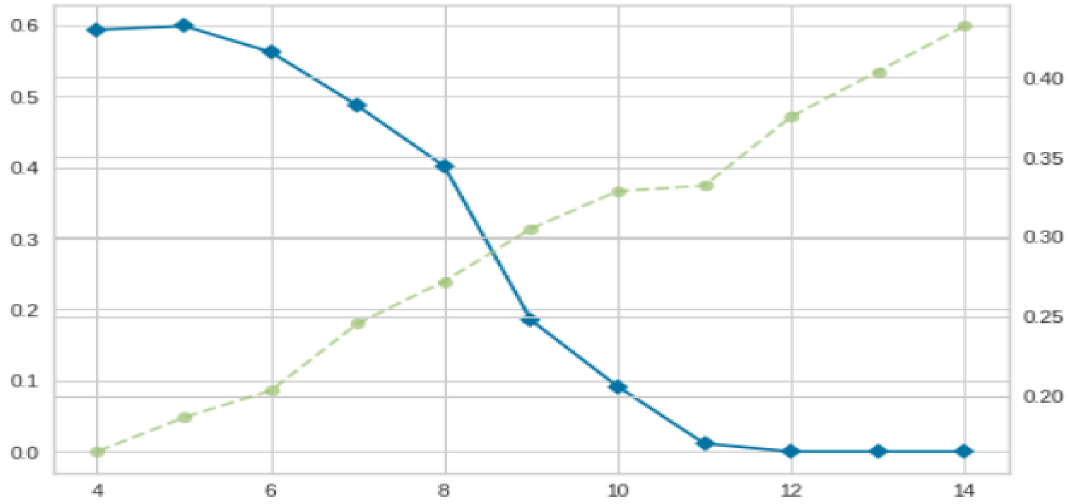


Figure 21 Elbow method clustering with eu_call embendings

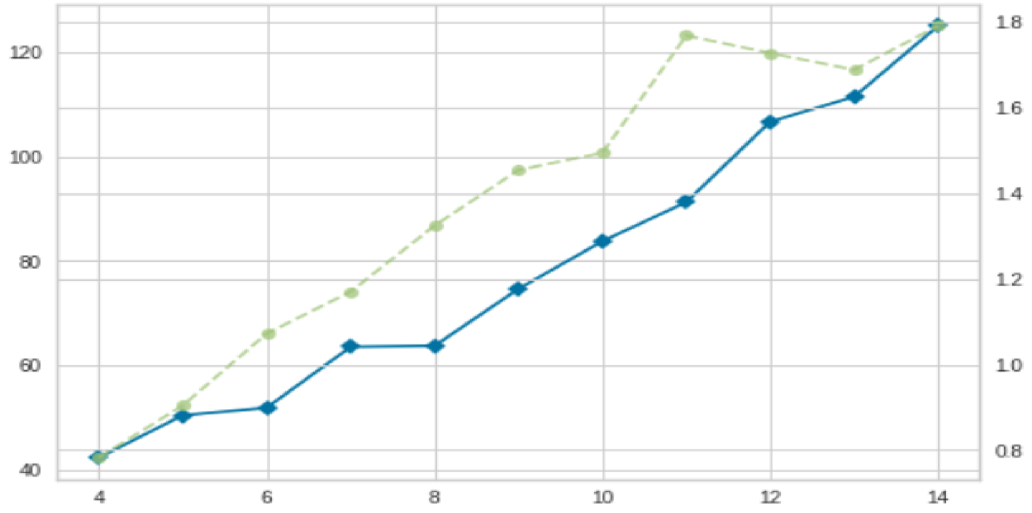


Figure 22 Elbow method with all document embeddings and claim

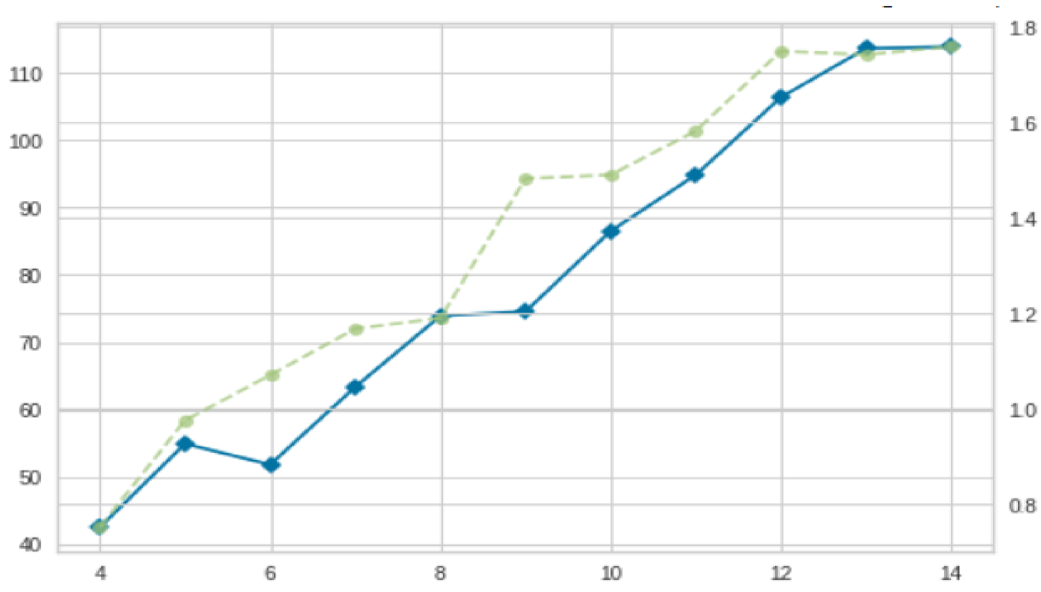


Figure 23 Elbow method with all document embeddings and claim evidence

3.3.1b Silhouette method

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point of one cluster is, related to the points of the neighboring clusters and thus provides a way to assess parameters, like number of clusters visually. This measure has a range of [-1, 1]. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters. Negative values indicate that those samples might have been assigned to the wrong cluster. The silhouette score falls within the range [-1, 1]. In other words, the silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect.

Our result is that all cases belong to the category that all abstracts' clusters are overlapping, thus it is difficult to organize scientific subject by examining their sentences.

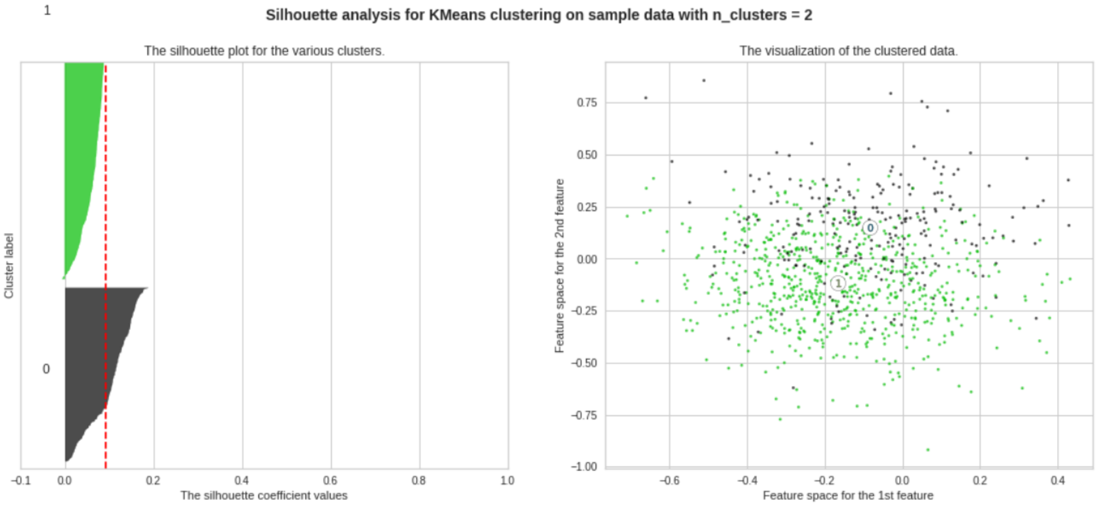


Figure 24 Silhouette method for abstract embeddings k=2

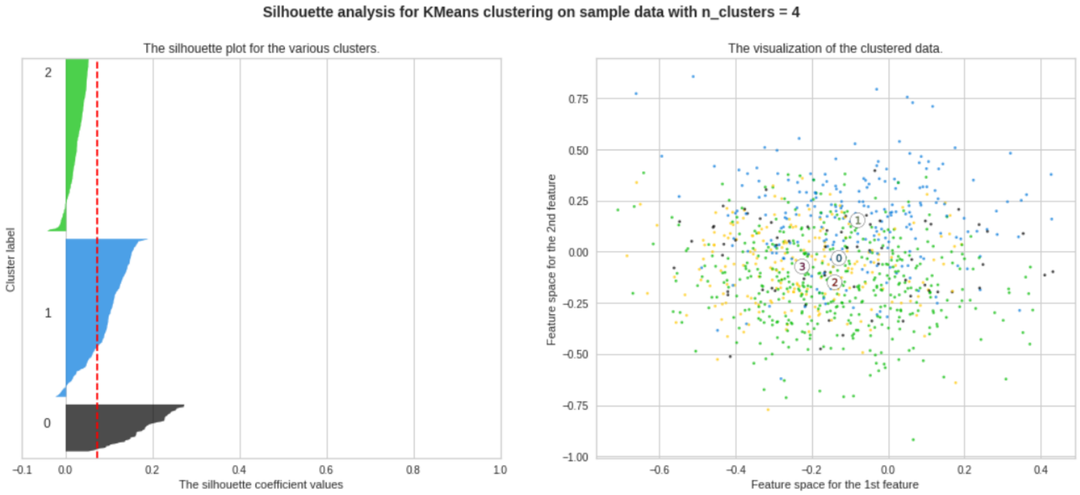


Figure 25 Silhouette method for abstract embeddings k=4

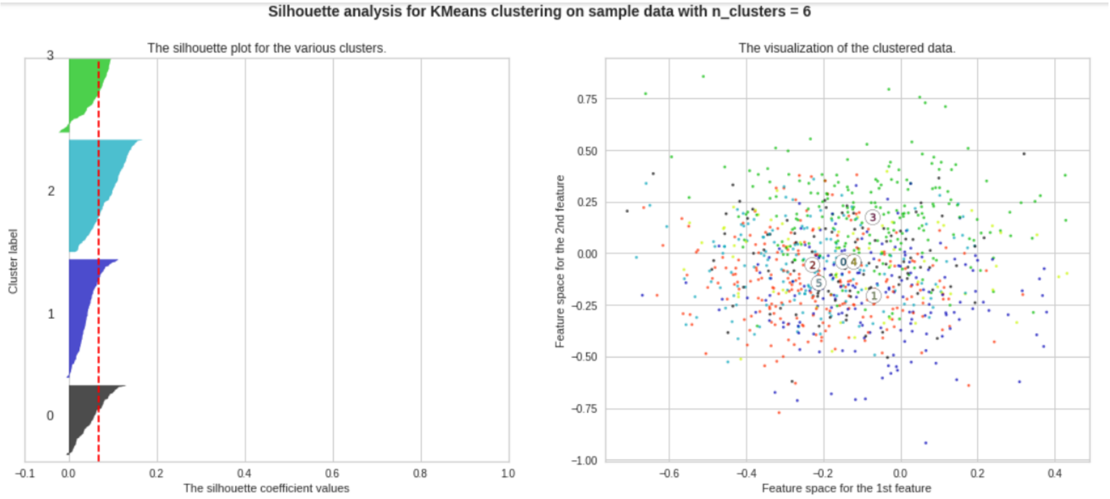


Figure 26 Silhouette method for abstract embeddings k=6

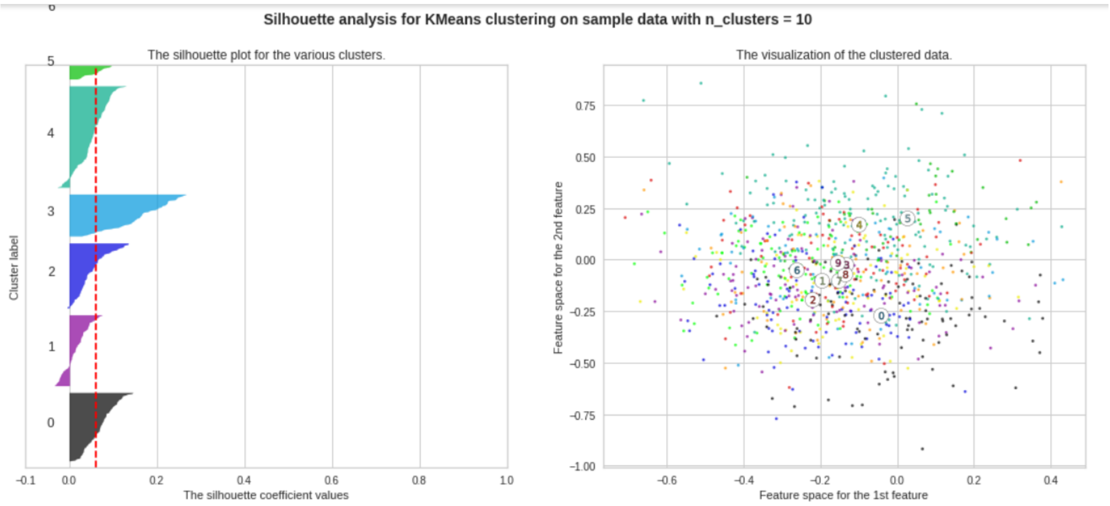
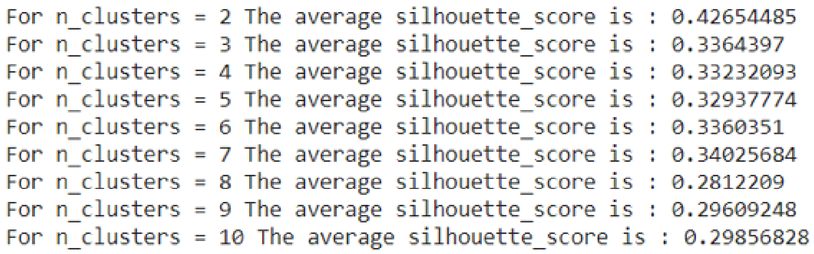


Figure 27 Silhouette method for abstract embeddings k=10

Then, we repeat the same process for the abstracts and features. Here, while the silhouette_score seems to be more acceptable (Figure 28) the clusters either are overlapping again or the size of clusters is totally different. As result, it is difficult again to examine the data.



```
For n_clusters = 2 The average silhouette_score is : 0.42654485
For n_clusters = 3 The average silhouette_score is : 0.3364397
For n_clusters = 4 The average silhouette_score is : 0.33232093
For n_clusters = 5 The average silhouette_score is : 0.32937774
For n_clusters = 6 The average silhouette_score is : 0.3360351
For n_clusters = 7 The average silhouette_score is : 0.34025684
For n_clusters = 8 The average silhouette_score is : 0.2812209
For n_clusters = 9 The average silhouette_score is : 0.29609248
For n_clusters = 10 The average silhouette_score is : 0.29856828
```

Figure 28 Silhouette Scores for the method for abstract embeddings and features

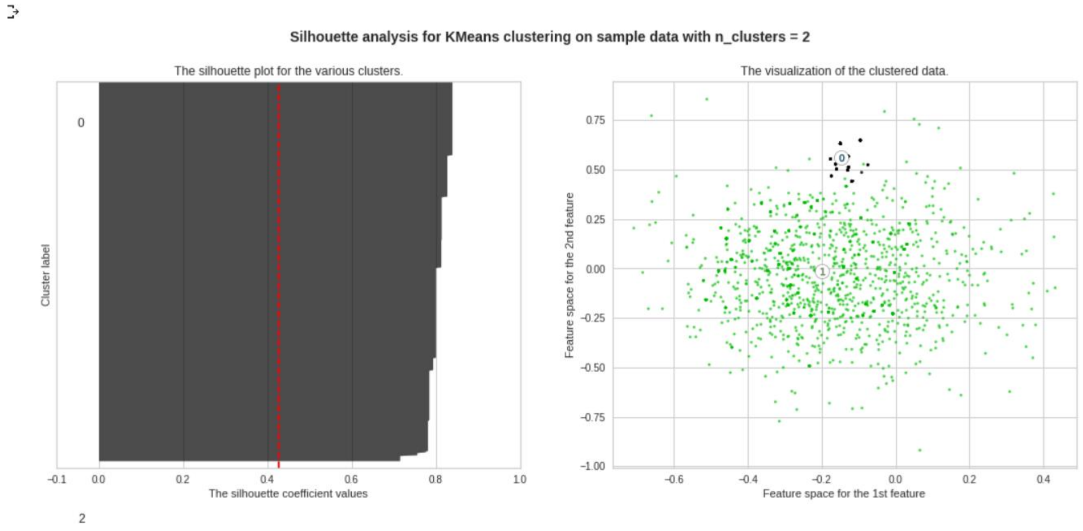


Figure 29 Silhouette method for abstract embeddings and features k=2

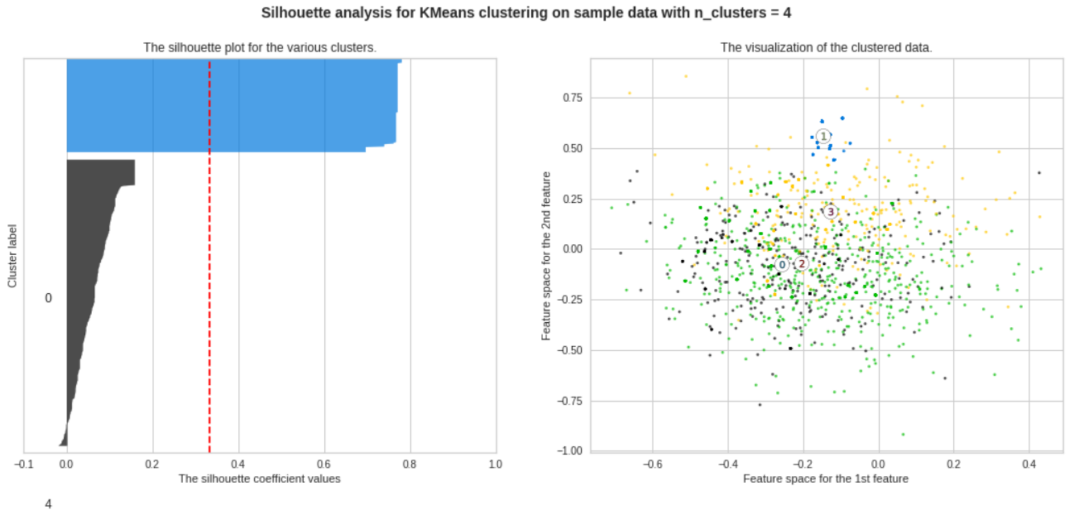


Figure 30 Silhouette method for abstract embeddings and features k=4

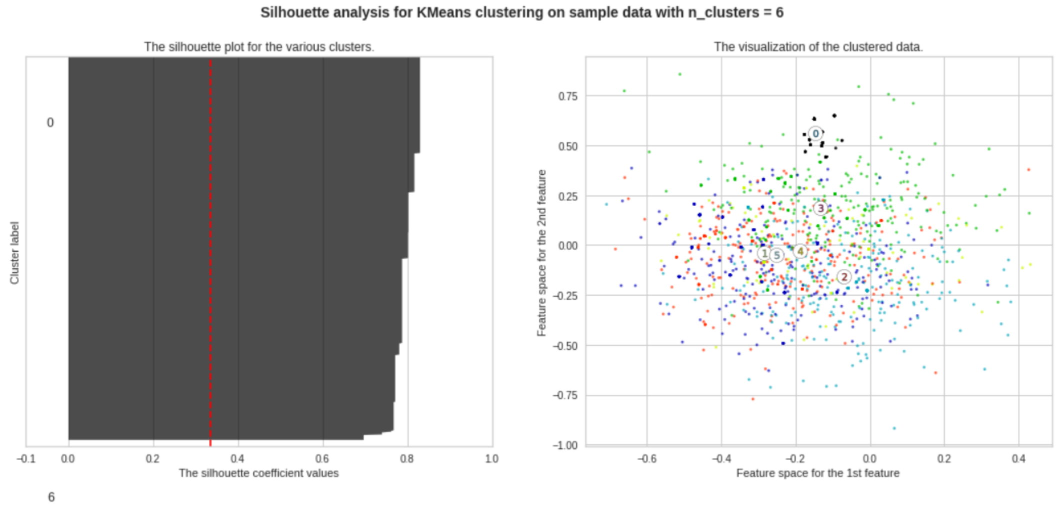


Figure 31 Silhouette method for abstract embeddings and features k=6

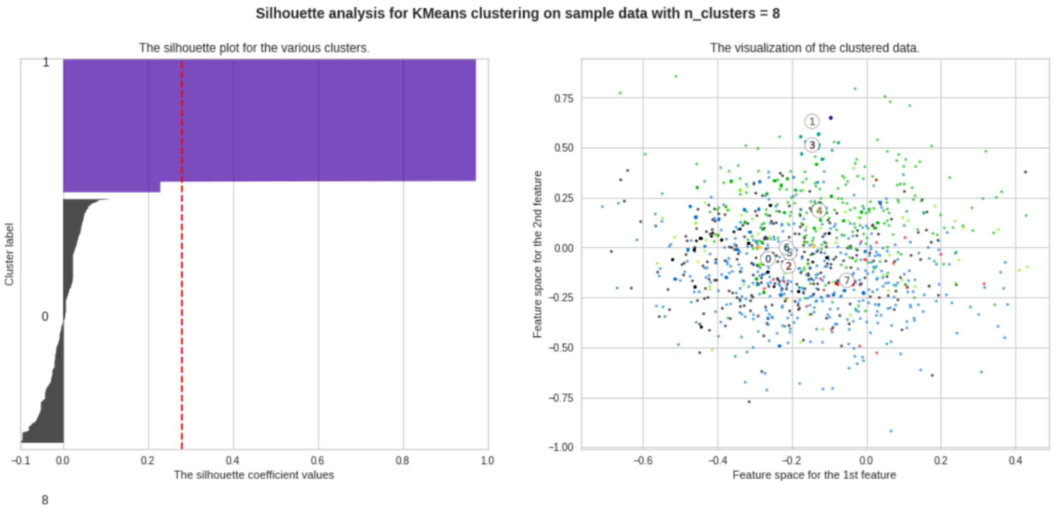


Figure 32 Silhouette method for abstract embeddings and features k=8

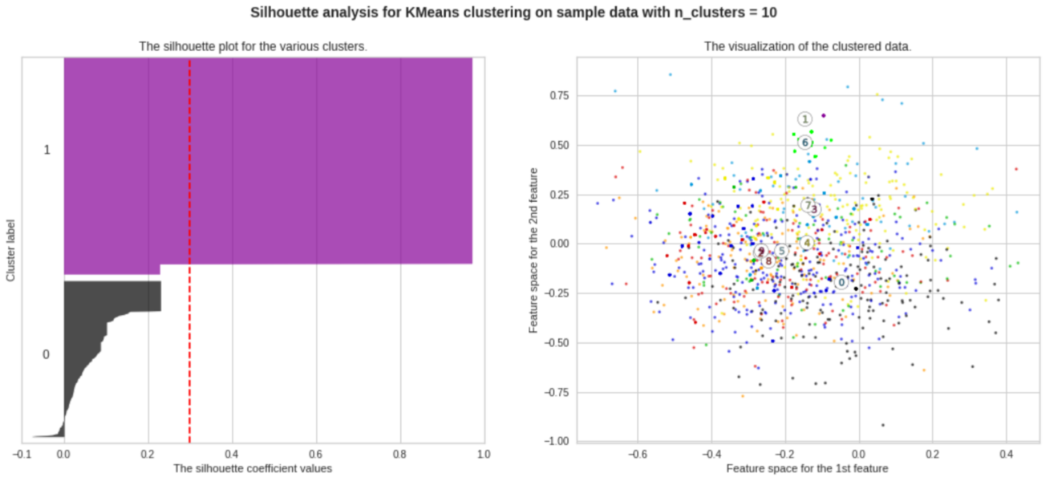


Figure 33 Silhouette method for abstract embeddings and features k=10

4. Results and Quantitave Analysis and Errors

4.1 Visualizations

All project’s visualizations have been respectively incorporated in previous sections.
(Figure 1- Figure 33)

4.2 Quantitave Analysis and Errors

As we show in Classification Report, the model for argument prediction could not predict well claim and evidence, in comparison to the neither label, the two first were predicted with errors. Perhaps, a reason that this event occurs, is because of the way the words were tokenized or combined with the previous and the words after. In other words, maybe there were key-phrases all over 105 abstracts that were repeated declaring claim or evidence and our model due to its performance on keywords could not increase his success rate. The same explanation we give for structure prediction as well. Regarding errors in Clustering, the common line in both methods was that using sentences' embeddings is difficult to cluster abstracts. A reason would be that scientific abstracts utilize common words to form paragraphs, structures, methodologies etc. For this reason, when clustering method is looking to organize a bag of words in groups, it is difficult to discern which word goes in the correspondent bin. This project gave us the chance to meet and understand the power of neural networks and clustering capabilities, but as always there is space for ameliorate a prediction's model.

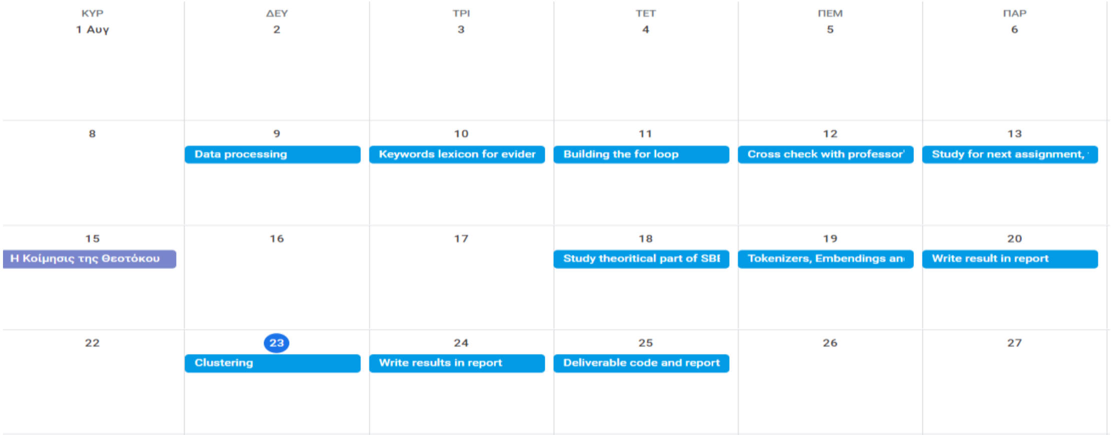
5. Discussion and Future goals

Machine learning journey was a great experience, from the annotation phase to the implementation of code and interpretation of result. Our feature goals as team certainly would be to investigate other algorithms of Machine learning and expand our knowledge in other kinds of recognition like image or voice.

6. Members and Roles

As we said in previous section, our team consists of three members all of them are women. All effort was deducted in team spirit, everyone did anything,which means there were not any role assigned. The result of this assignment is a collective effort, aiming to return knowledge and useful information to his members and the scientific community as well.

7. Time Plan



ΚΥΡ	ΔΕΥ	ΤΡΙ	ΤΕΤ	ΠΕΜ	ΠΑΡ
1 Αυγ	2	3	4	5	6
8	9 Data processing	10 Keywords lexicon for evider	11 Building the for loop	12 Cross check with professor'	13 Study for next assignment, '
15 Η Κοίμησις της Θεοτόκου	16	17	18 Study theoritical part of SBI	19 Tokenizers, Embendings an	20 Write result in report
22	23 Clustering	24 Write results in report	25 Deliverable code and report	26	27

Figure 3x Time Plan

8. Table of figures

Figure 1 Data Collection in the Label Studio 3
Figure 2 Argument dataset 4
Figure 3 Structure dataset 4
Figure 4 Citations dataset 5
Figure 5 Cross check Argument Label Claim 6
Figure 6 Cross check Argument Label Evidence 7
Figure 7 Feed Forward Network Schema 8
Figure 8 Count the frequency of top words 9
Figure 9 Fit argument prediction model over 32 epochs 10
Figure 10 Model metric loss 11
Figure 11 Model metric accuracy 11
Figure 12 Loss-Accuracy interpration 12
Figure 13 Confusion Matrix Argeument Prediction evaluation 12
Figure 14 Classification Report Argument Prediction evaluation 12
Figure 15 Fit structure prediction model over 32 epochs 14
Figure 16 Loss metric structure prediction 15
Figure 17 Accuracy structure prediction 15
Figure 18 Confusion Matrix for Structure Prediction 15
Figure 19 Classification Report for structure prediction 16
Figure 20 Elbow method for abstract embendings 17
Figure 21 Elbow method clustering with eu_call embendings 17
Figure 22 Elbow method with all document embeddings and claim 17
Figure 23 Elbow method with all document embdendings and claim_evidence 18
Figure 24 Silhouette method for abstract embeddings k=2 19
Figure 25 Silhouette method for abstract embeddings k=4 19
Figure 26 Silhouette method for abstract embeddings k=6 20
Figure 27 Silhouette method for abstract embeddings k=10 20
Figure 28 Silhouette Scores for the method for abstract embeddings and features 208
Figure 29 Silhouette method for abstract embeddings and features k=2 209
Figure 30 Silhouette method for abstract embeddings and features k=4 209
Figure 31 Silhouette method for abstract embeddings and features k=6 20
Figure 32 Silhouette method for abstract embeddings and features k=8 20
Figure 32 Silhouette method for abstract embeddings and features k=10 20

9. Bibliography

Websites

https://en.wikipedia.org/wiki/Feedforward_neural_network
https://hal.inria.fr/hal-02448197v2/document
https://stackoverflow.com/questions/50184280/how-to-conceptually-think-about-relationship-between-tokenized-words-and-word-em
https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
https://dylancastillo.co/nlp-snippets-cluster-documents-using-word2vec/#cluster-documents-using-mini-batches-k-means
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
https://towardsdatascience.com/stratified-splitting-of-grouped-datasets-using-optimization-bdc12fb6e691
https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526
https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/
https://www.datarobot.com/wiki/fitting/
https://www.baeldung.com/cs/ml-loss-accuracy
https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397
https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam