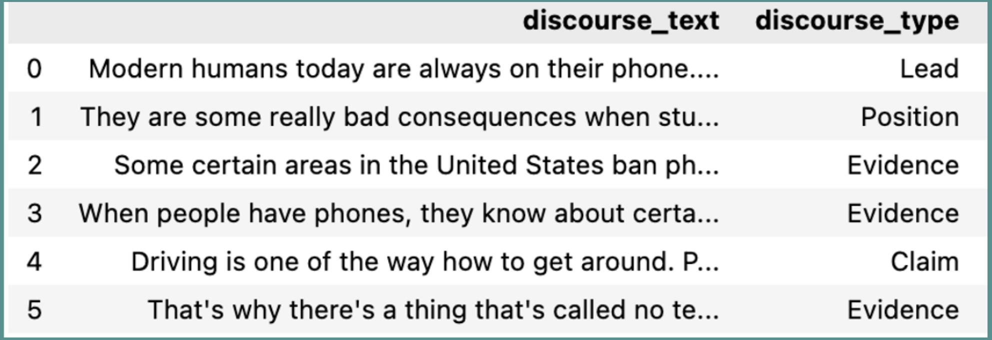
NLP Project: Student Writing Evaluation

- Discourse Elements Classification

Group 1: Ruiqi Li, Yixi Liang, He Huang, Yuan Dang

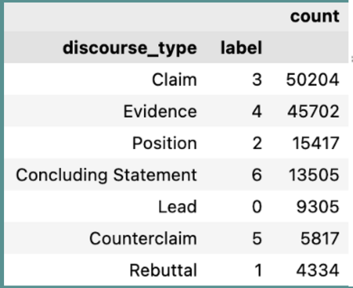
Dataset

- **Source** – Kaggle Competition(https://www.kaggle.com/competitions/feedback-prize-2021/overview)
- **Introduction -** The dataset contains argumentative essays written by U.S students in grades 6-12, they are automatically segmented into discrete discourse elements.
- **Goal -** classify discourse elements
- **Shape** – 144,280 x 8
  - id - ID code for essay response
  - discourse_id - ID code for discourse element
  - discourse_start - character position where discourse element begins in the essay response
  - discourse_end - character position where discourse element ends in the essay response
  - **discourse_text - text of discourse element**
  - **discourse_type - (target) classification of discourse element**
  - discourse_type_num - enumerated class label of discourse element
  - predictionstring - the word indices of the training sample, as required for predictions

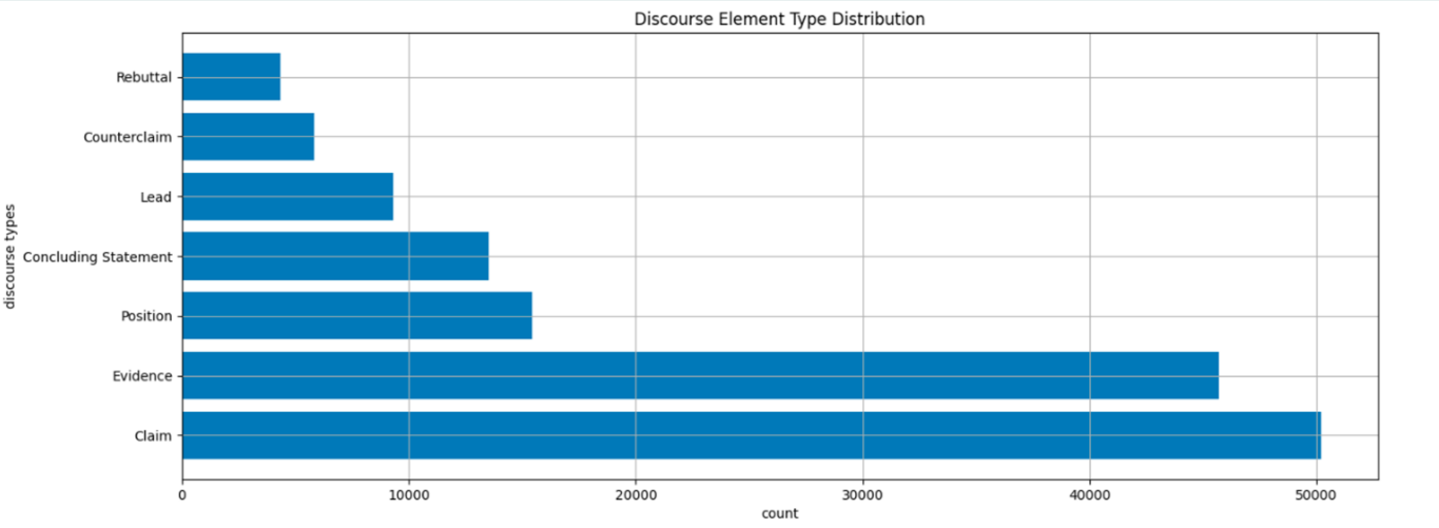


	discourse_text	discourse_type
0	Modern humans today are always on their phone....	Lead
1	They are some really bad consequences when stu...	Position
2	Some certain areas in the United States ban ph...	Evidence
3	When people have phones, they know about certa...	Evidence
4	Driving is one of the way how to get around. P...	Claim
5	That's why there's a thing that's called no te...	Evidence

Dataset



		count
discourse_type	label	
Claim	3	50204
Evidence	4	45702
Position	2	15417
Concluding Statement	6	13505
Lead	0	9305
Counterclaim	5	5817
Rebuttal	1	4334

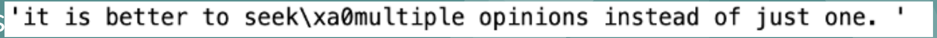


Target – 7 categories

- **Lead** - an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader’s attention and point toward the thesis
- **Position** - an opinion or conclusion on the main question
- **Claim** - a claim that supports the position
- **Counterclaim** - a claim that refutes another claim or gives an opposing reason to the position
- **Rebuttal** - a claim that refutes a counterclaim
- **Evidence** - ideas or examples that support claims, counterclaims, or rebuttals.
- **Concluding Statement** - a concluding statement that restates the claims

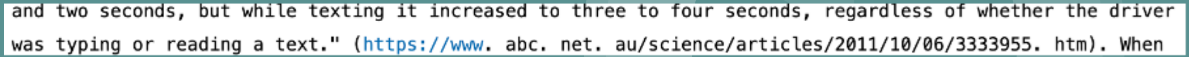
Clean Text

- Remove non-ASCII characters



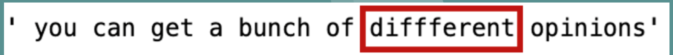
```
'it is better to seek\xa0multiple opinions instead of just one. '
```

- Remove url
- Remove parenthesis



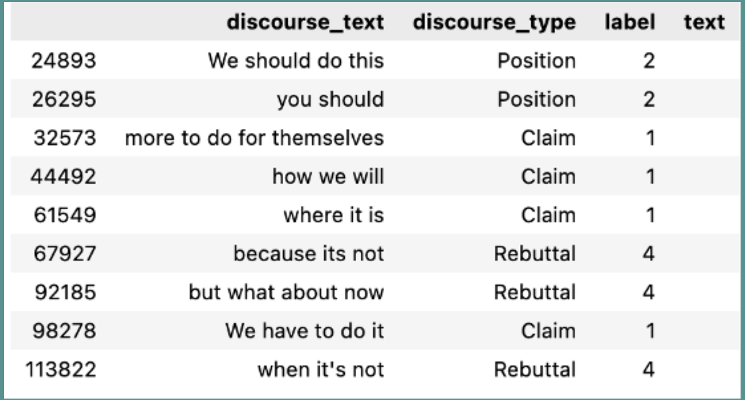
```
and two seconds, but while texting it increased to three to four seconds, regardless of whether the driver
was typing or reading a text." (https://www. abc. net. au/science/articles/2011/10/06/3333955. htm). When
```

- Expand contraction
- Fix meaningless repeating characters



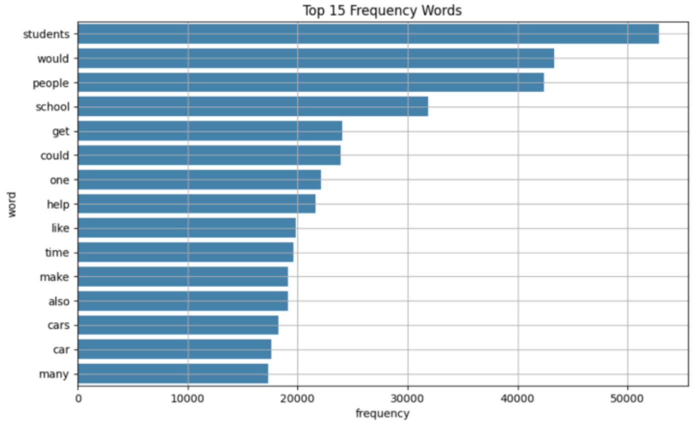
```
' you can get a bunch of diffferent opinions'
```

- Remove NLTK stop words
- Remove meaningless text



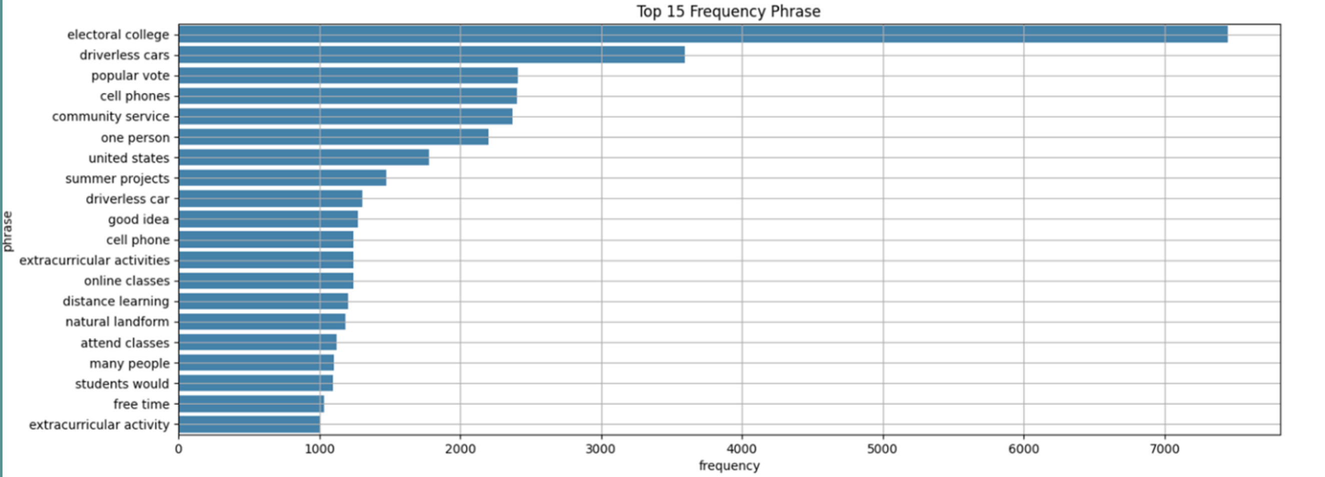
	discourse_text	discourse_type	label	text
24893	We should do this	Position	2	
26295	you should	Position	2	
32573	more to do for themselves	Claim	1	
44492	how we will	Claim	1	
61549	where it is	Claim	1	
67927	because its not	Rebuttal	4	
92185	but what about now	Rebuttal	4	
98278	We have to do it	Claim	1	
113822	when it's not	Rebuttal	4	

Words & Phrases Extraction


Top 15 Frequency Words

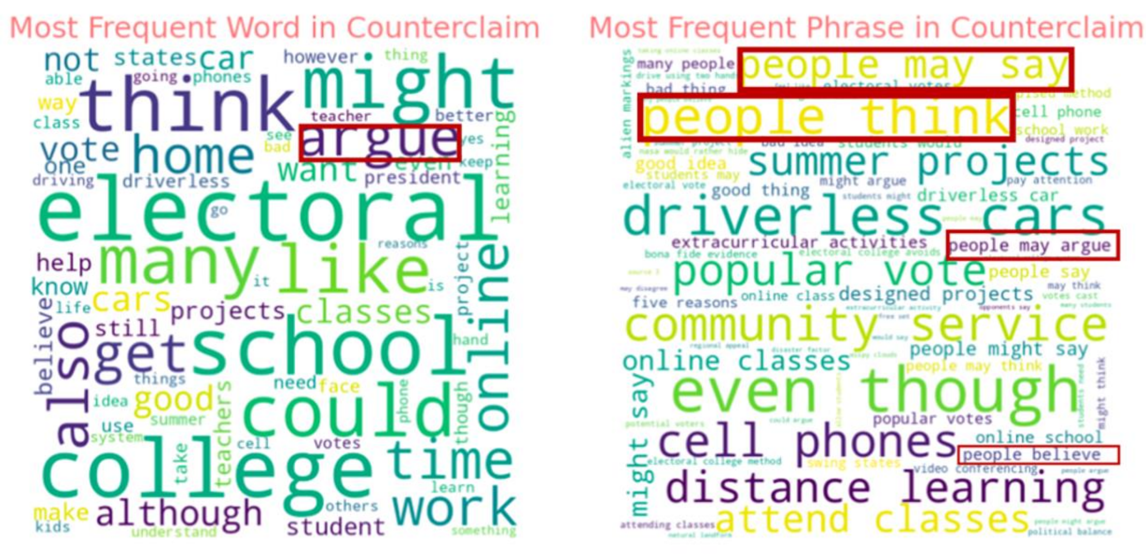
Most Common Topics:

- Student, School, Extracurricular activity
- Electoral college, Vote
- Driverless cars


Top 15 Frequency Phrase

WordCloud

Word Cloud of counterclaim



Most Frequent Word in Counterclaim

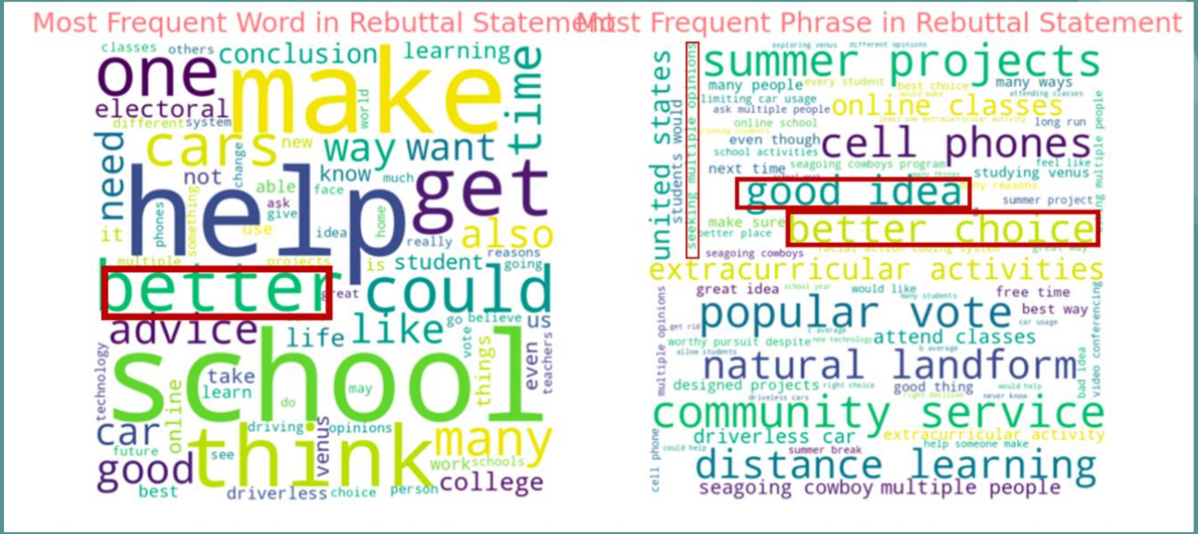
Most Frequent Phrase in Counterclaim

Counterclaim

- presents an opposite argument to the author’s argument
- always starts with “people may argue” .etc
- keywords ‘argue’

WordCloud

Word Cloud of rebuttal



Most Frequent Word in Rebuttal Statement

Most Frequent Phrase in Rebuttal Statement

Rebuttal Statement

- refutes counterclaim to strengthen the author's argument
- keywords
  - 'good idea',
  - 'better choice',
  - 'seeking multiple opinions'

Rule-based models

- Preprocessed Data (Tokenization, Stem, Stopwords Removal, TFIDF, etc)
- Into Naive Bayes Classifier
- Into Logistic Regression
- Into Logistic Regression (in LSA form)

Naive Bayes

- Bayes Rule
- posterior probability of an event given the probability of another event that has happened
- All features are independent of each other

- A & B are events, P(B) must not be 0

- P(A|B) : posterior probability of A given B.
- P(A): prior probability of A
- P(B|A): likelihood probability of B given A
- P(B): prior probability of B

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Logistic

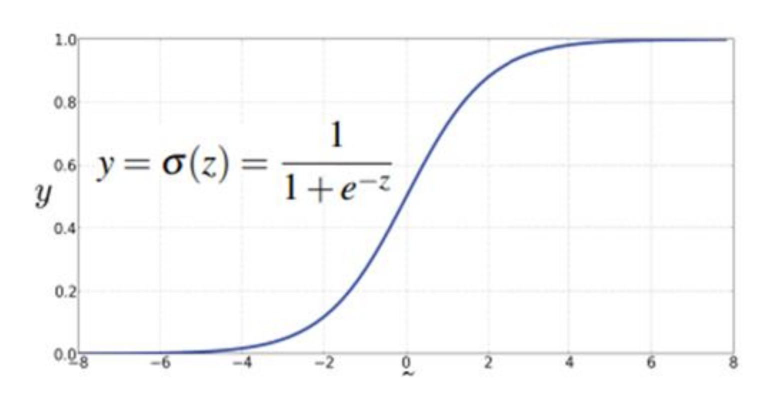
- Baseline supervised ML for classification

- Input x: [x1, x1, ..., xn]
- Weights W: [w1, w2, ..., wn]
- Z: Sum up all the weighted features and the bias

- y: a function of z that goes from 0 to 1 (sigmoid)

$$Z = (\sum_{i=1}^{n} w_i x_i) + b$$

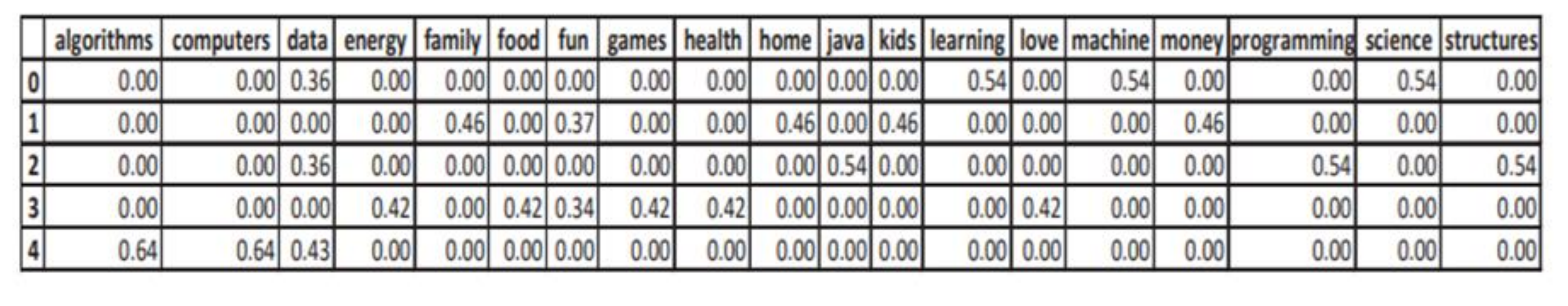
$$Z = w \cdot x + b$$



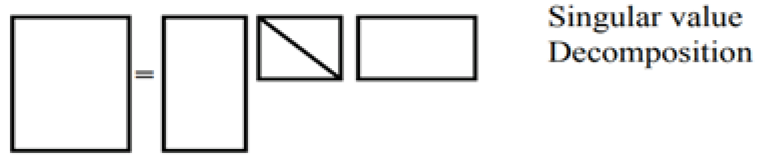
$$\hat{y} = 1 if P(y = 1|x) > 0.5 otherwise 0$$

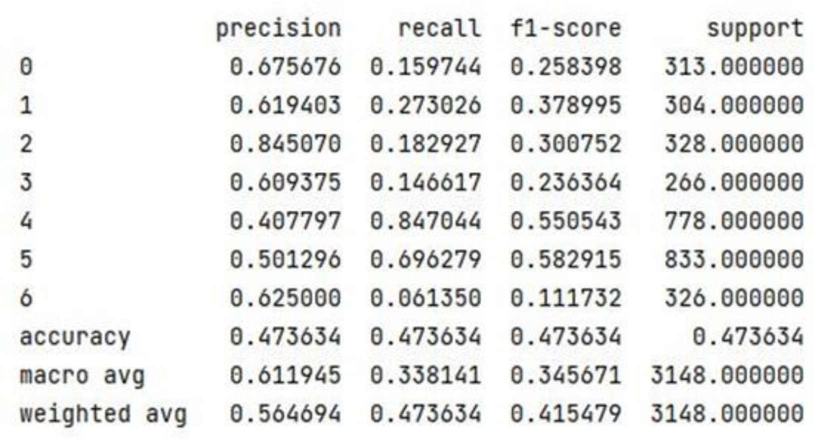
About LSA

- Word -> score of importance in each array
- Arrays -> lower-dimensional data
- To improve computational efficiency



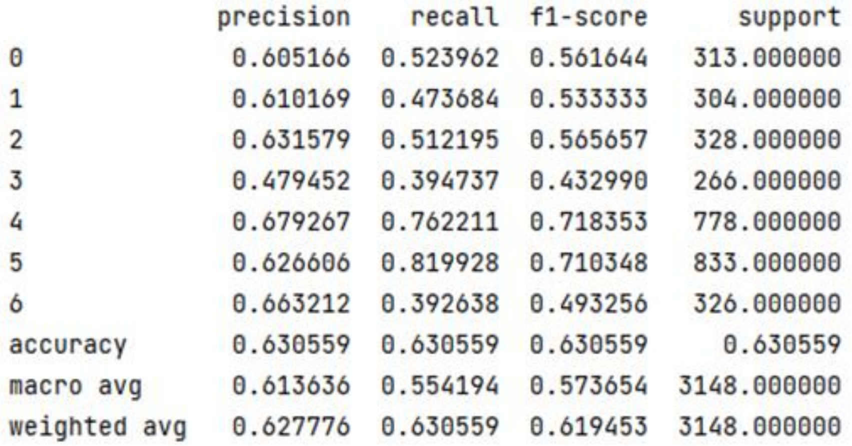
	algorithms	computers	data	energy	family	food	fun	games	health	home	java	kids	learning	love	machine	money	programming	science	structures
0	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.54	0.00	0.00	0.54	0.00
1	0.00	0.00	0.00	0.00	0.46	0.00	0.37	0.00	0.00	0.46	0.00	0.46	0.00	0.00	0.00	0.46	0.00	0.00	0.00
2	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.54
3	0.00	0.00	0.00	0.42	0.00	0.42	0.34	0.42	0.42	0.00	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.00	0.00
4	0.64	0.64	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00





```
              precision    recall  f1-score      support
0              0.675676  0.159744  0.258398   313.000000
1              0.619403  0.273026  0.378995   304.000000
2              0.845070  0.182927  0.300752   328.000000
3              0.609375  0.146617  0.236364   266.000000
4              0.407797  0.847044  0.550543   778.000000
5              0.501296  0.696279  0.582915   833.000000
6              0.625000  0.061350  0.111732   326.000000
accuracy       0.473634  0.473634  0.473634     0.473634
macro avg      0.611945  0.338141  0.345671  3148.000000
weighted avg   0.564694  0.473634  0.415479  3148.000000
```

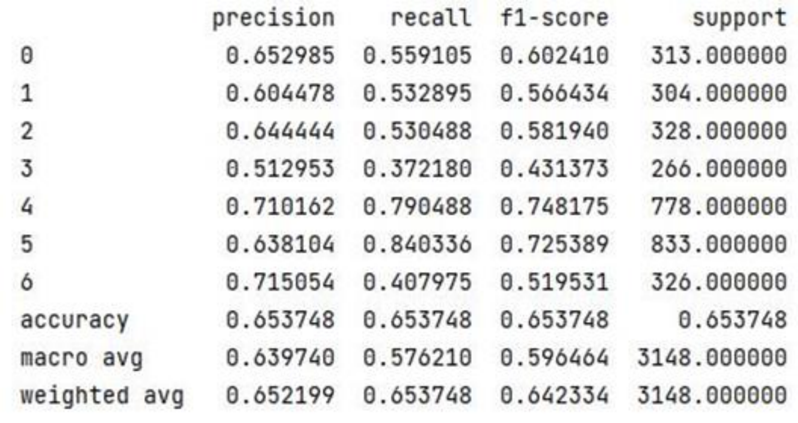
Naive Bayes Report



```
              precision    recall  f1-score      support
0              0.605166  0.523962  0.561644   313.000000
1              0.610169  0.473684  0.533333   304.000000
2              0.631579  0.512195  0.565657   328.000000
3              0.479452  0.394737  0.432990   266.000000
4              0.679267  0.762211  0.718353   778.000000
5              0.626606  0.819928  0.710348   833.000000
6              0.663212  0.392638  0.493256   326.000000
accuracy       0.630559  0.630559  0.630559     0.630559
macro avg      0.613636  0.554194  0.573654  3148.000000
weighted avg   0.627776  0.630559  0.619453  3148.000000
```

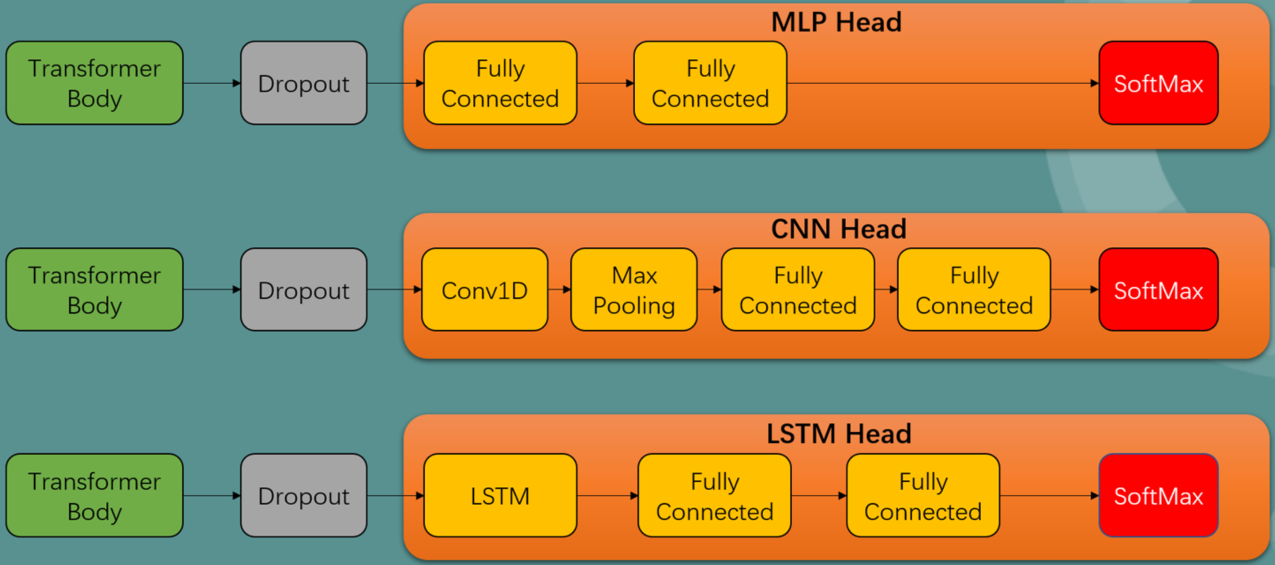
LSA Logistic Report (n_components = 500, n_iters = 1000)

Logistic Regression Report



```
              precision    recall  f1-score      support
0              0.652985  0.559105  0.602410   313.000000
1              0.604478  0.532895  0.566434   304.000000
2              0.644444  0.530488  0.581940   328.000000
3              0.512953  0.372180  0.431373   266.000000
4              0.710162  0.790488  0.748175   778.000000
5              0.638104  0.840336  0.725389   833.000000
6              0.715054  0.407975  0.519531   326.000000
accuracy       0.653748  0.653748  0.653748     0.653748
macro avg      0.639740  0.576210  0.596464  3148.000000
weighted avg   0.652199  0.653748  0.642334  3148.000000
```

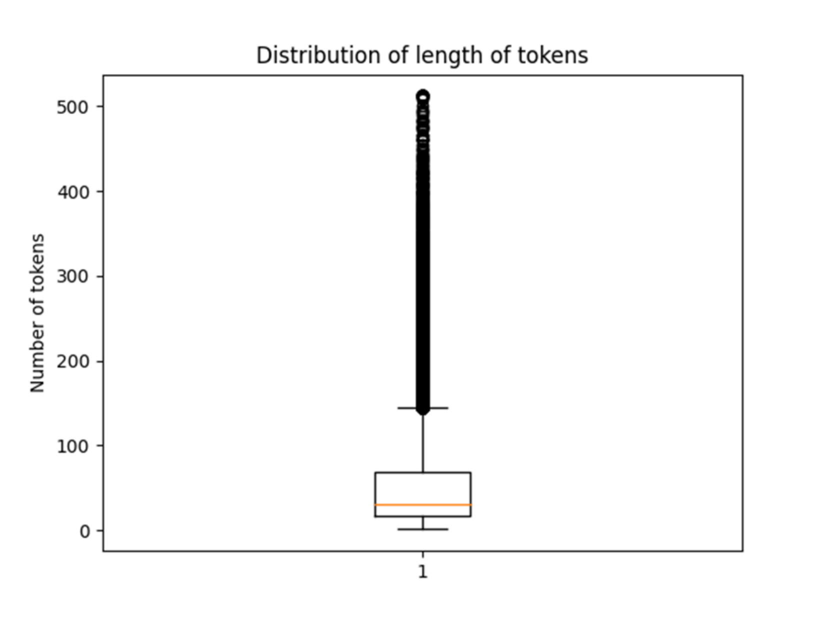
Transformer



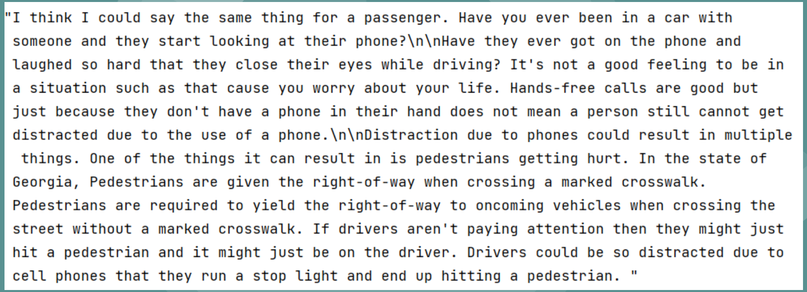
Transformer:

BERT, RoBerta, XLM-Roberta

Transformer



Tokenizer example: BERT

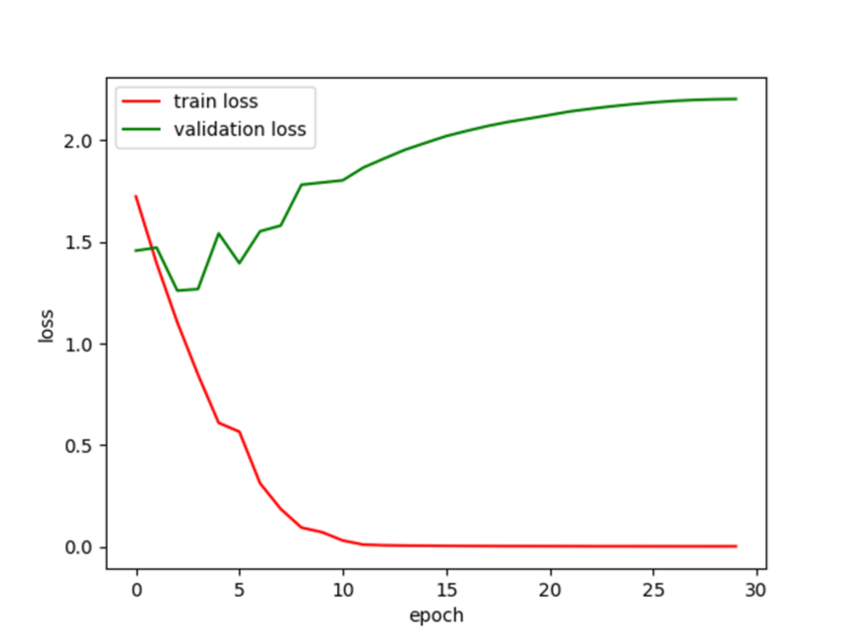


```
"I think I could say the same thing for a passenger. Have you ever been in a car with
 someone and they start looking at their phone?\n\nHave they ever got on the phone and
 laughed so hard that they close their eyes while driving? It's not a good feeling to be in
 a situation such as that cause you worry about your life. Hands-free calls are good but
 just because they don't have a phone in their hand does not mean a person still cannot get
 distracted due to the use of a phone.\n\nDistraction due to phones could result in multiple
  things. One of the things it can result in is pedestrians getting hurt. In the state of
 Georgia, Pedestrians are given the right-of-way when crossing a marked crosswalk.
 Pedestrians are required to yield the right-of-way to oncoming vehicles when crossing the
 street without a marked crosswalk. If drivers aren't paying attention then they might just
 hit a pedestrian and it might just be on the driver. Drivers could be so distracted due to
 cell phones that they run a stop light and end up hitting a pedestrian. "
```

Outlier example

Set Max length : 150

Transformer



Overfitting

Add dropout layers

Improve dropout rate

Reduce learning rate

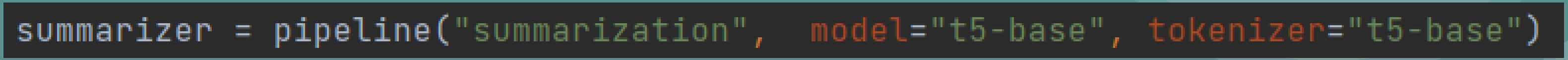
Transformer

Transformer body	Custom Head	Accuracy	F1-macro
bert-base-cased	MLP	0.759847522	0.733070241
bert-base-cased	CNN	0.764612452	0.729890715
bert-base-cased	LSTM	0.783989835	0.760746121
xlm-roberta-base	LSTM	0.643348561	0.620621515

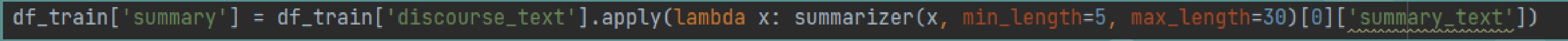
Differences between transformers are bigger than differences between custom heads

Summary

Mean text length	Max text length
45.336898962084305	589



```
summarizer = pipeline("summarization",  model="t5-base", tokenizer="t5-base")
```



```
df_train['summary'] = df_train['discourse_text'].apply(lambda x: summarizer(x, min_length=5, max_length=30)[0]['summary_text'])
```

Summary

Before summarize	After summarize
If teachers give their students coming into their classes a summer break project that the teacher created to help introduce them to the topic they will be learning it will make understanding the subject easier.	if teachers give their students coming into their classes a summer break project they created it will make understanding the subject easier .
In Bogota, Colombia they had a care-free day and it turned into a big hit. On this car-free day millions of Colombians hiked, biked, skated, or took buses. Leaving the streets of this capital city eerily deviod of traffic jams. You know but not everyone was willing to parcisepte in this event and those people faced twenty-five dollar fines. This program was set to sprend to other countries and it did. For the first time, two other Colombian cities, Cali and Valledupar joined the event, and they were happy.	millions of Colombians hiked, biked, skated, or took buses . the program was set to sprend to other countries

Summary

Before summarize	After summarize
The United States has a very large amount of unhealthy people. People who don't get to walk or do fitness regularly to be able to stay healthy and fit. Having limited the use of your car it would as help you mentally. In the article "In Germany Suburb, Life Goes On Without Cars" by Elisabeth Rosenthal Heidrun Walter a mother of two says " When i had a car i was always tense. I'm much happier this way." If a mother of two says that not using her car makes her happier. In the article "Car-free day is spinning into a big hit in Bogota." by Andrew selsky a businessman said " It's a good way to take away stress and lower air pollution." So if you ask me the would would be better without using cars everyday it's a win win situation.	the u.s. has a very large amount of unhealthy people . people who don't get to walk or do fitness regularly

Summary

Pretrained	Model	Accuracy	F1 macro
bert-base-cased	BERT+LSTM	0.7820838627700127	0.7575876668982763
bert-base-uncased	BERT+LSTM	0.7916137229987293	0.7685482955944065

Conclusion

In this final project, we preprocess the data and generate EDA visualizations to help us understand the data at the beginning.

Firstly, we build rule-based models of naïve bayes and logistic models. Logistic mode performs better than naïve bayes. LSA does not quite enhance the logistic model. The input data's dimensionality may not need a necessary reduction.

Secondly, of all the transformer body plus custom head models, BERT-LSTM model is the best one. The difference between different transformers body is larger than the difference between different custom head.

Finally, we try to use summarization to improve the result of the model. But there are still some problems remain. For instance, we do not know whether add original text and summary together is right or not to train the model. In addition, the summary of the texts is sometime not good enough. Moreover, it takes plenty of time to use summary transformer.

Thank you