Style Change Detection using BERT

Aarish Iyer and Soroush Vosoughi
Department of Computer Science, Dartmouth College,
Hanover, NH 03755

Aarish Iyer, and Soroush Vosoughi (2020). Style Change Detection Using BERT. In CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org.

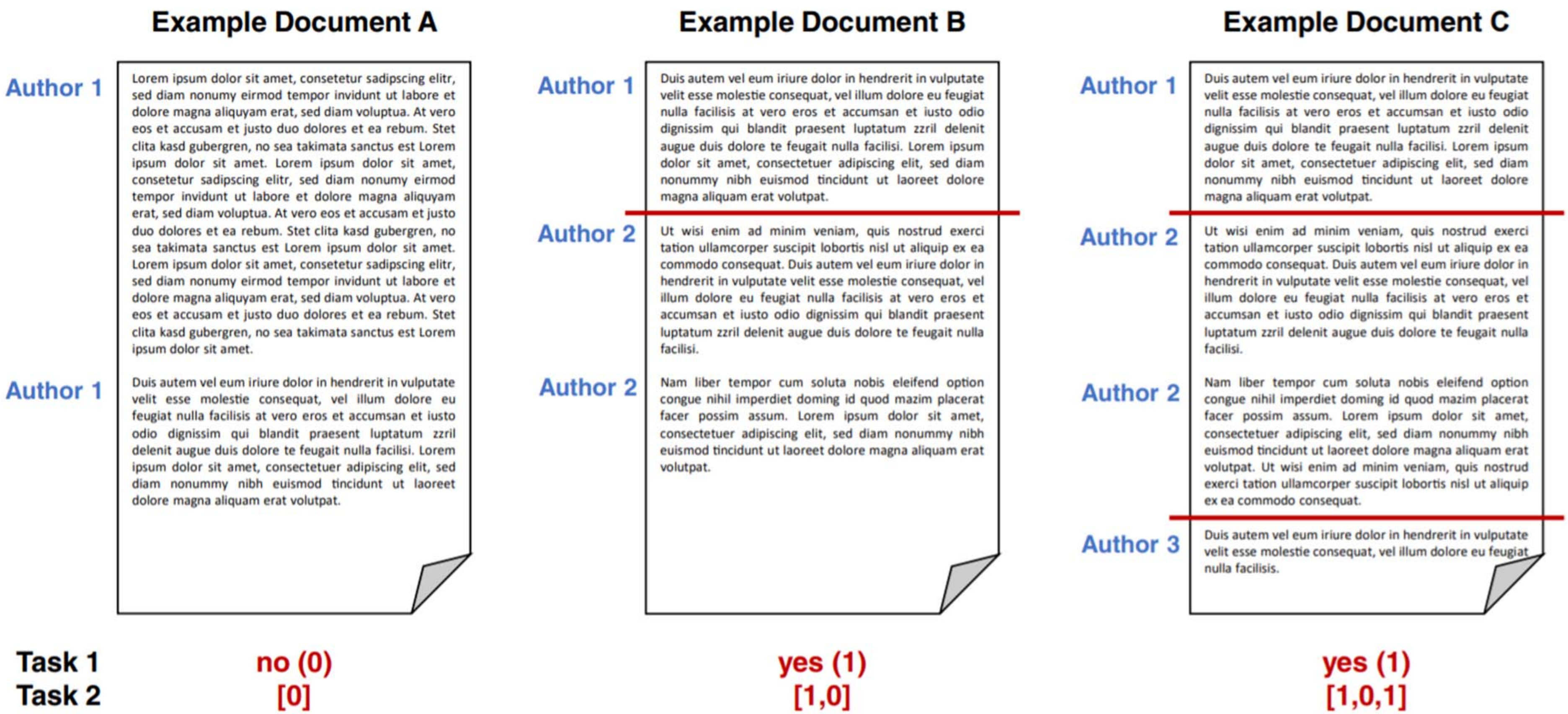
Task

This research was submitted as a solution to the Style Change Detection Challenge held by PAN@CLEF.

There were two sub-tasks for the challenge:

1. Given a document, is the document written by multiple authors?

2. Given a sequence of paragraphs of a (supposedly) multi-author document, is there a style change between any of the paragraphs?



Eva Zangerle, Maximilian Mayerl, Günther Specht, Martin Potthast, Benno Stein (2020). Overview of the Style Change Detection Task at PAN 2020. In CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org.

DataSet

- All the data was extracted from the StackExchange family of websites

DataSet

- All the data was extracted from the StackExchange family of websites

- There were two datasets provided for the task:
    - `Dataset-narrow` : Questions and answers from a specific subset of StackExchange sites pertaining to topics of Computer Technology.

	Narrow
Train	3,442
Validation	1,722

Table 1: Number of documents in each dataset

DataSet

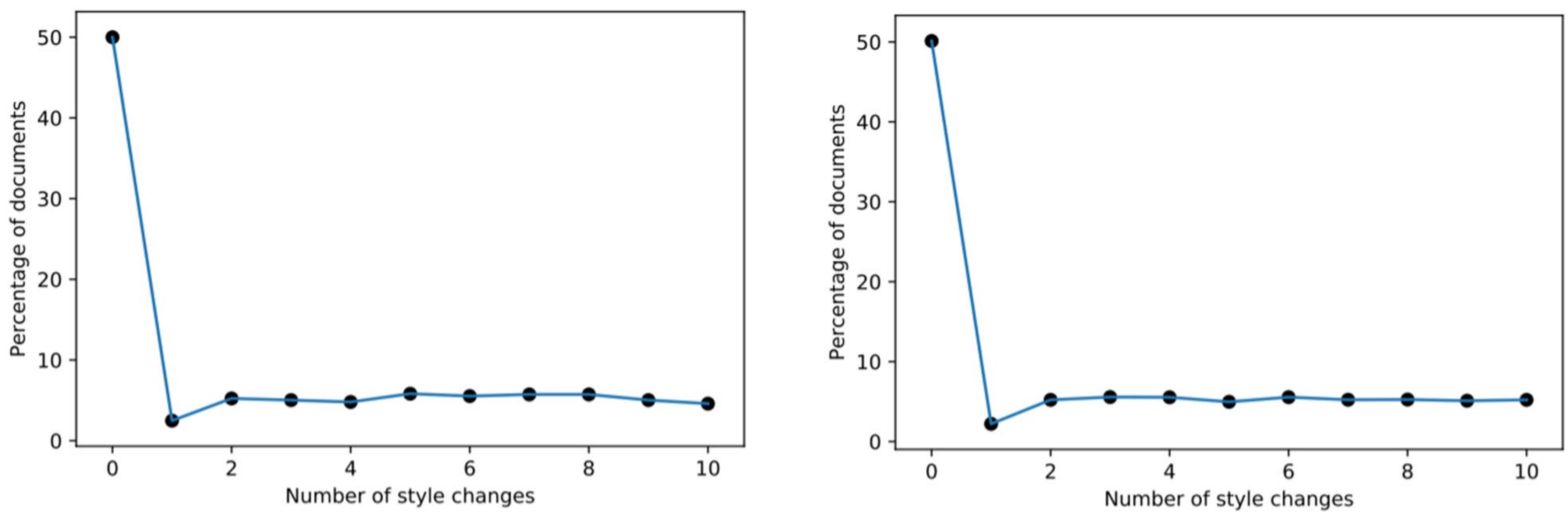
- All the data was extracted from the StackExchange family of websites

- There were two datasets provided for the task:
  - `Dataset-narrow` : Questions and answers from a specific subset of StackExchange sites pertaining to topics of Computer Technology.
  - `Dataset-wide` : Questions and answers from a subset of StackExchange sites that pertained to a wide variety of topics (Technology, Economics, Literature, Philosophy, and Mathematics).

	Narrow	Wide
Train	3,442	8,138
Validation	1,722	4,078

Table 1: Number of documents in each dataset

DataSet



(a) Narrow (b) Wide

Figure 1: Distribution of number of style changes in different datasets

Bidirectional Encoder Representations from Transformers (BERT)

BERT is a large-scale pre-trained deep model used for solving a variety of NLP tasks, obtaining state-of-the-art results on various benchmarks.

Of all the BERT models available, the BERT Base Cased model was used (layers= 12, hidden size= 768, self-attention heads= 12, total parameters= 110M).

Bidirectional Encoder Representations from Transformers (BERT)

BERT is a large-scale pre-trained deep model used for solving a variety of NLP tasks, obtaining state-of-the-art results on various benchmarks.

Of all the BERT models available, the BERT Base Cased model was used (layers= 12, hidden size= 768, self-attention heads= 12, total parameters= 110M)

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Approach

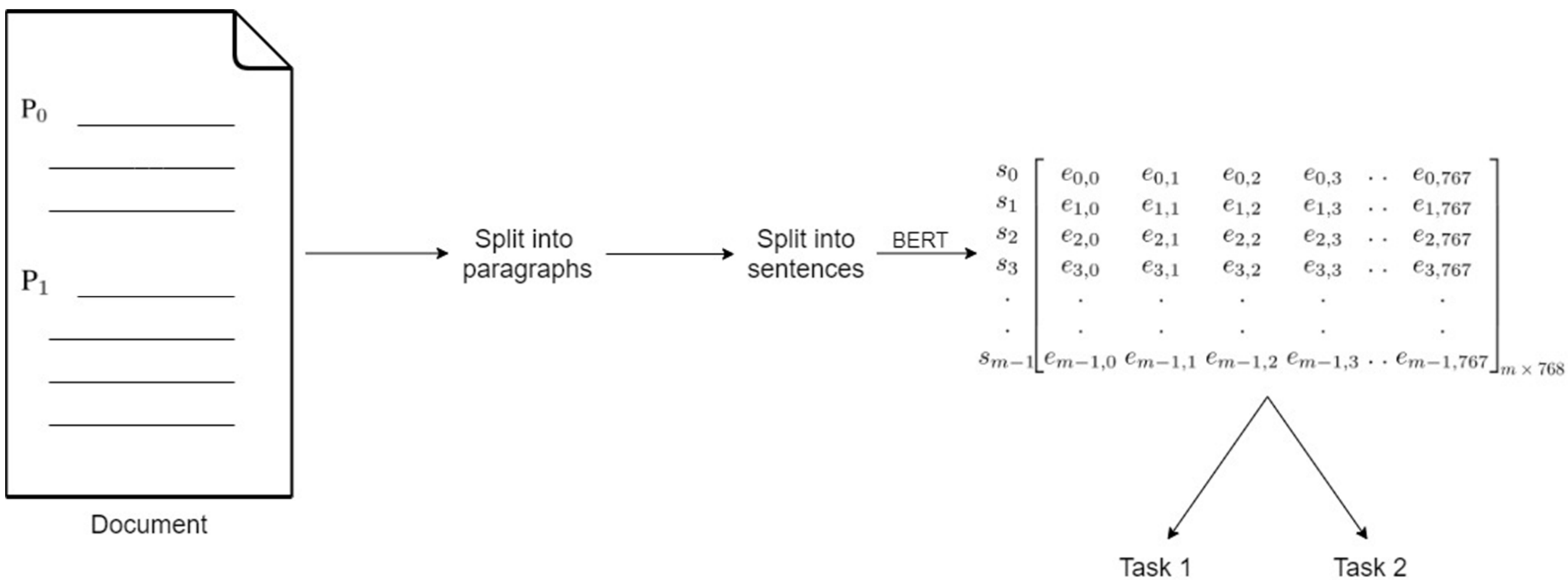
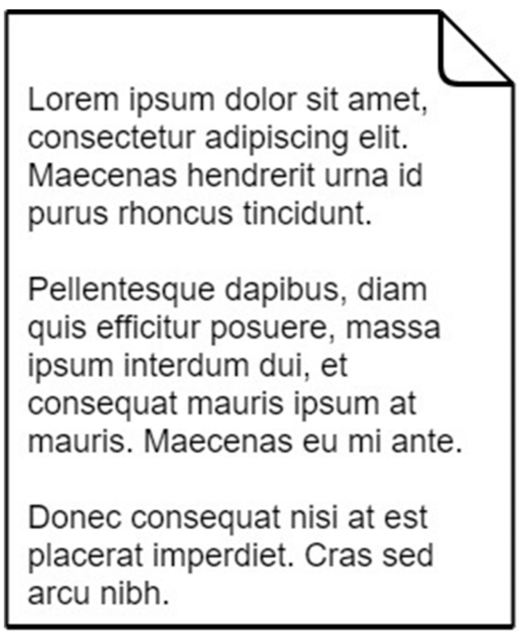


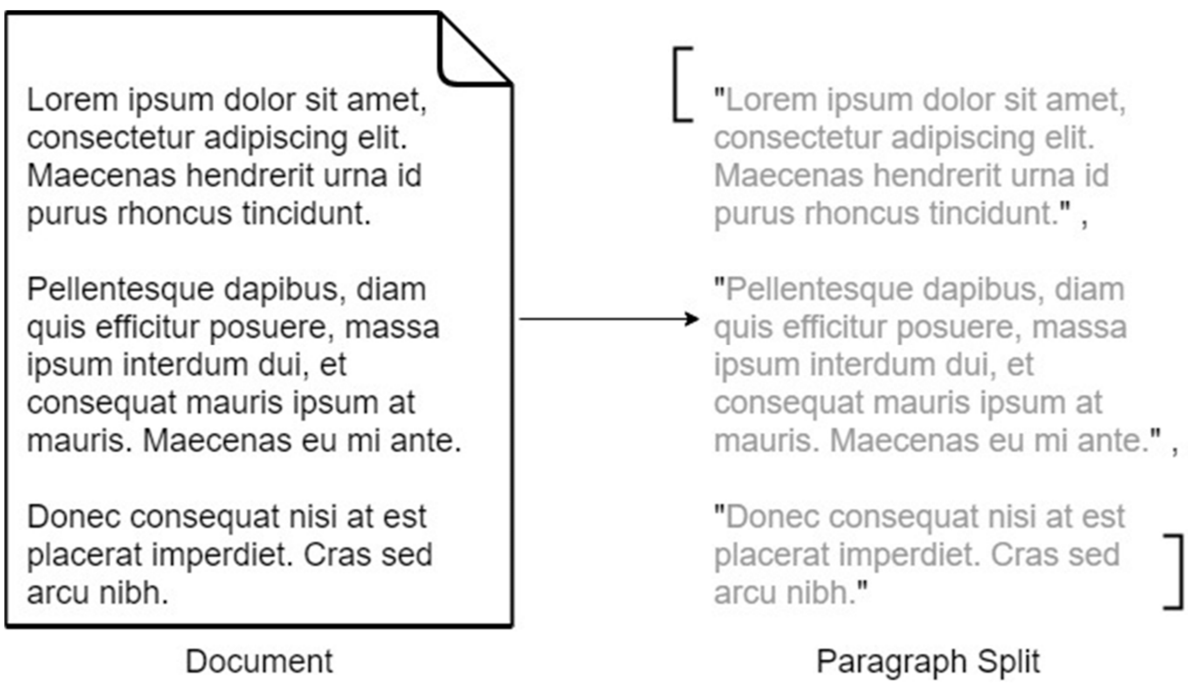
Figure 3: Our approach for generating feature vectors for the two tasks using pretrained BERT

Approach

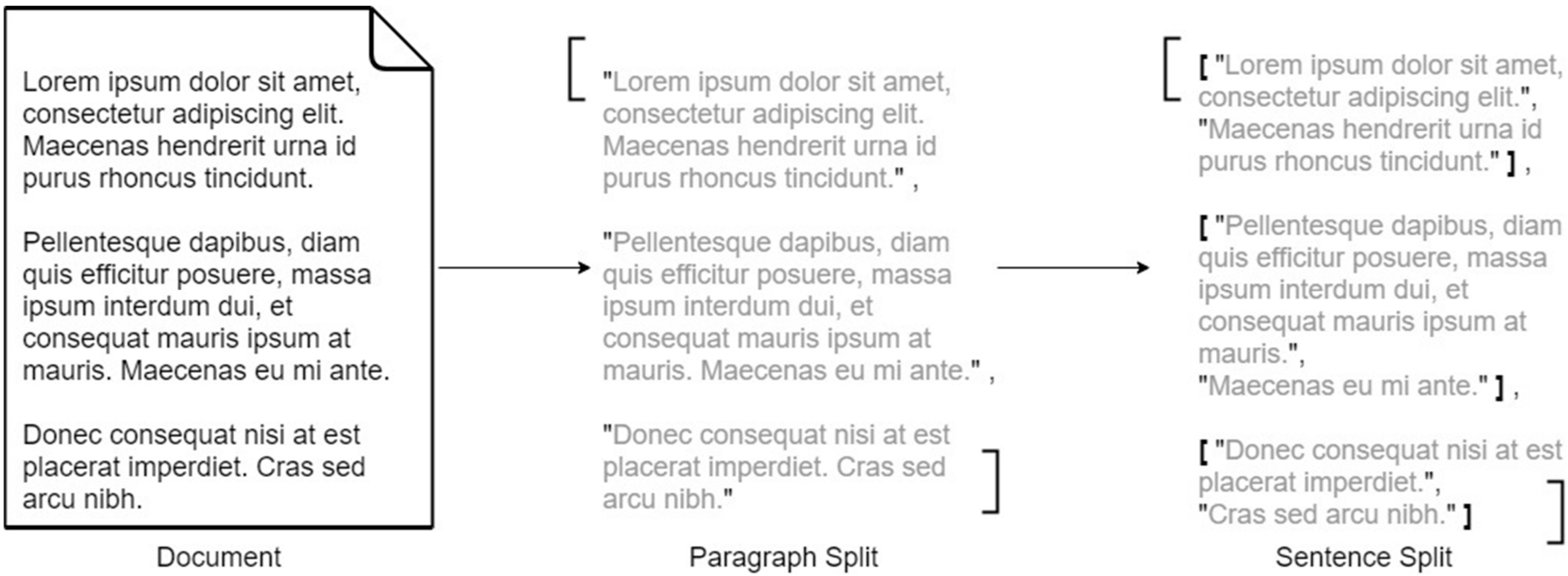


Document

Approach


Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas hendrerit urna id purus rhoncus tincidunt.
Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris. Maecenas eu mi ante.
Donec consequat nisi at est placerat imperdiet. Cras sed arcu nibh.
Document
[
"Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas hendrerit urna id purus rhoncus tincidunt." ,
"Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris. Maecenas eu mi ante." ,
"Donec consequat nisi at est placerat imperdiet. Cras sed arcu nibh."
]
Paragraph Split

Approach


Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas hendrerit urna id purus rhoncus tincidunt.
Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris. Maecenas eu mi ante.
Donec consequat nisi at est placerat imperdiet. Cras sed arcu nibh.
Document
["Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas hendrerit urna id purus rhoncus tincidunt." ,
"Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris. Maecenas eu mi ante." ,
"Donec consequat nisi at est placerat imperdiet. Cras sed arcu nibh."]
Paragraph Split
[["Lorem ipsum dolor sit amet, consectetur adipiscing elit.",
"Maecenas hendrerit urna id purus rhoncus tincidunt."] ,
["Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris.",
"Maecenas eu mi ante."] ,
["Donec consequat nisi at est placerat imperdiet.",
"Cras sed arcu nibh."]]
Sentence Split

Approach

["Lorem ipsum dolor sit amet, consectetur adipiscing elit.", "Maecenas hendrerit urna id purus rhoncus tincidunt."] ,

["Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris.", "Maecenas eu mi ante."] ,

["Donec consequat nisi at est placerat imperdiet.", "Cras sed arcu nibh."]

Sentence Split

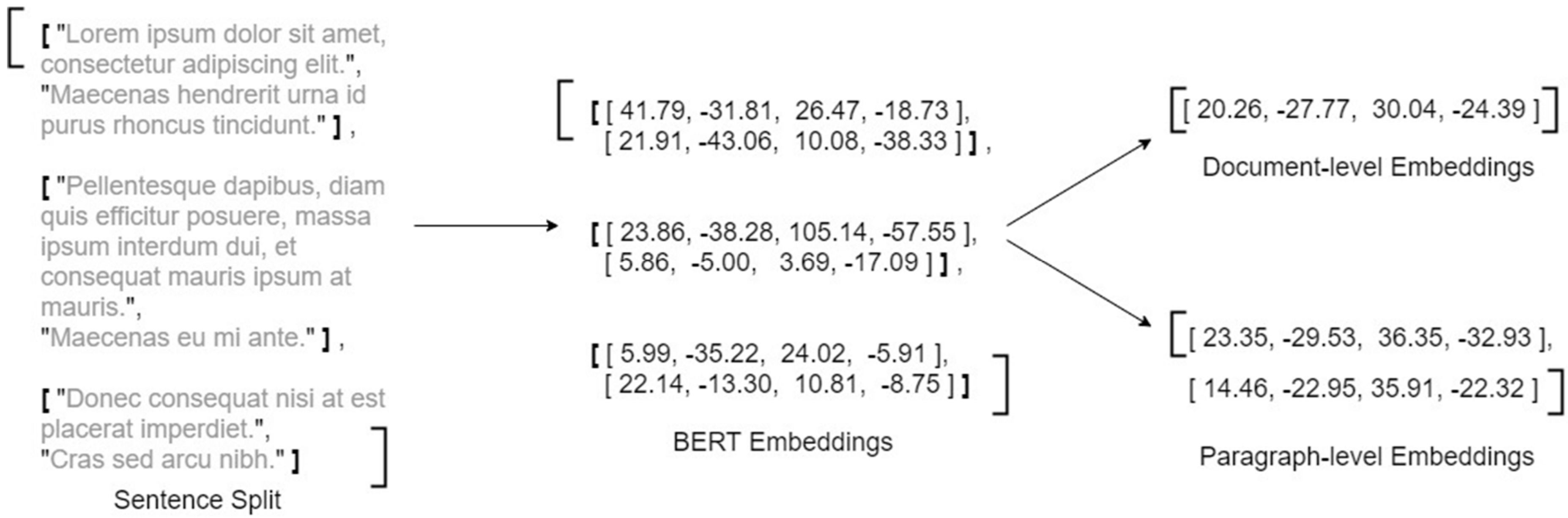
[[41.79, -31.81, 26.47, -18.73],
[21.91, -43.06, 10.08, -38.33]] ,

[[23.86, -38.28, 105.14, -57.55],
[5.86, -5.00, 3.69, -17.09]] ,

[[5.99, -35.22, 24.02, -5.91],
[22.14, -13.30, 10.81, -8.75]]

BERT Embeddings

Approach


["Lorem ipsum dolor sit amet, consectetur adipiscing elit.", "Maecenas hendrerit urna id purus rhoncus tincidunt."] ,
["Pellentesque dapibus, diam quis efficitur posuere, massa ipsum interdum dui, et consequat mauris ipsum at mauris.", "Maecenas eu mi ante."] ,
["Donec consequat nisi at est placerat imperdiet.", "Cras sed arcu nibh."]
Sentence Split
[[41.79, -31.81, 26.47, -18.73], [21.91, -43.06, 10.08, -38.33]] ,
[[23.86, -38.28, 105.14, -57.55], [5.86, -5.00, 3.69, -17.09]] ,
[[5.99, -35.22, 24.02, -5.91], [22.14, -13.30, 10.81, -8.75]]
BERT Embeddings
[20.26, -27.77, 30.04, -24.39]
Document-level Embeddings
[23.35, -29.53, 36.35, -32.93],
[14.46, -22.95, 35.91, -22.32]
Paragraph-level Embeddings

Classifier

We tried various binary classifiers for Task 1 on `Dataset-wide.` The results obtained on the validation set are:

Classifier	**F-1 Score**
SVM	0.6504
Decision Tree	0.6108
Logistic Regression	0.6533
Gaussian Naive Baye’s	0.566
Random Forest	**0.7367**

Results

	Narrow	Wide
Document-level	0.7661	0.7575
Paragraph-level	0.8805	0.8306

Table 2: F1 scores calculated on the validation set for Document-level (task 1) and Paragraph-level (task 2) predictions.

	Average
Document-level	0.6401
Paragraph-level	0.8566

Table 3: Average F1 scores calculated on the test set for Document-level (task 1) and Paragraph-level (task 2) predictions

Other Methods

Creating a Dataset of sentence pairs: Each data point was a pair of sentences from consecutive paragraphs.

Other Methods

Creating a Dataset of sentence pairs: Each data point was a pair of sentences from consecutive paragraphs.

The label of the data point would be assigned based on the following policy:

- If the two sentences are from the same paragraph → 0
- If the two sentences are from different paragraphs
  - If no style change occurred between the two paragraphs → 0
  - If a style change occurred between the two paragraphs → 1

Other Methods

Creating a Dataset of sentence pairs: Each data point was a pair of sentences from consecutive paragraphs.

The label of the data point would be assigned based on the following policy:

- If the two sentences are from the same paragraph → 0
- If the two sentences are from different paragraphs
  - If no style change occurred between the two paragraphs → 0
  - If a style change occurred between the two paragraphs → 1

The dataset was severely imbalanced at this stage, so it was balanced by removing data points from the majority class at random.

Other Methods

The following methods were tried on the newly constructed sentence-pair Dataset

Other Methods

The following methods were tried on the newly constructed sentence-pair Dataset

Fine-tuning BERT:

- Fine-tune BERT using the sentence-pair dataset, and then perform the classification
- Accuracy plateaued after a point

Other Methods

The following methods were tried on the newly constructed sentence-pair Dataset

Fine-tuning BERT:

- Fine-tune BERT using the sentence-pair dataset, and then perform the classification
- Accuracy plateaued after a point

Convolutional Neural Network:

- The data points were converted to tensors of size $(l1 + l2) \times 768$
- Then run through kernels of sizes (2×768), (3×768), ..., (5×768)
- Experiments are ongoing with this technique

Pitfalls

Some of the disadvantages of our method are:

- **Runtime**
  - All experiments were run in an environment that had access to a GPU
  - Running on the validation set for `Dataset-wide` took about 2-3 hours

Pitfalls

Some of the disadvantages of our method are:

- **Runtime**
  - All experiments were run in an environment that had access to a GPU
  - Running on the validation set for `Dataset-wide` took about 2-3 hours

- **Only focuses on semantic features**
  - We believe that the best approach for style change detection would be to combine both semantic and stylistic features, but our method only focuses on semantic features for now.

Future Work

- **Fine-tuning BERT**
  - Since we only tried fine-tuning it with our custom dataset, it would be interesting to see the results by fine-tuning it with the original dataset

Future Work

- **Fine-tuning BERT**
  - Since we only tried fine-tuning it with our custom dataset, it would be interesting to see the results by fine-tuning it with the original dataset

- **Combining Semantic and Syntactic features**
  - A more sophisticated approach which takes into consideration both Semantic and Stylistic features would be the next step to improve the current model.

THANK YOU