Advanced Machine Learning

Final project proposal

by
Tomer Segal, ID301833833
Millis Sahar, ID300420379

Instructions:
The project proposal should address the following points.
The length is no longer than 2 pages.

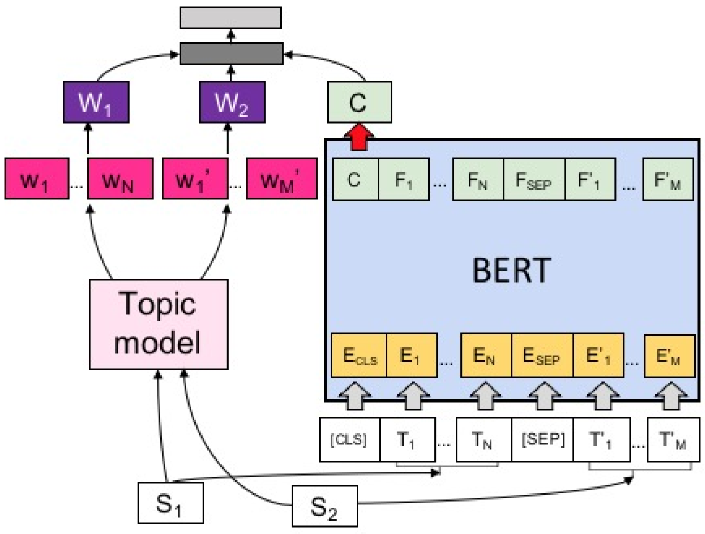
Anchor paper

State the anchor paper

Our research will be based on the article "tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection" and the its code.

Provide a very short summary of the paper

They propose a novel topic-informed BERT-based architecture for pairwise **semantic similarity** detection and show that their model improves performance over strong neural baselines across a variety of English language datasets.
They also find that the addition of topics to BERT helps particularly with resolving domain-specific cases.


W1
W2
C
w1 ... wN
w1' ... wM'
C F1 ... FN FSEP F'1 ... F'M
BERT
Topic model
ECLS E1 ... EN ESEP E'1 ... E'M
[CLS] T1 ... TN [SEP] T'1 ... T'M
S1
S2

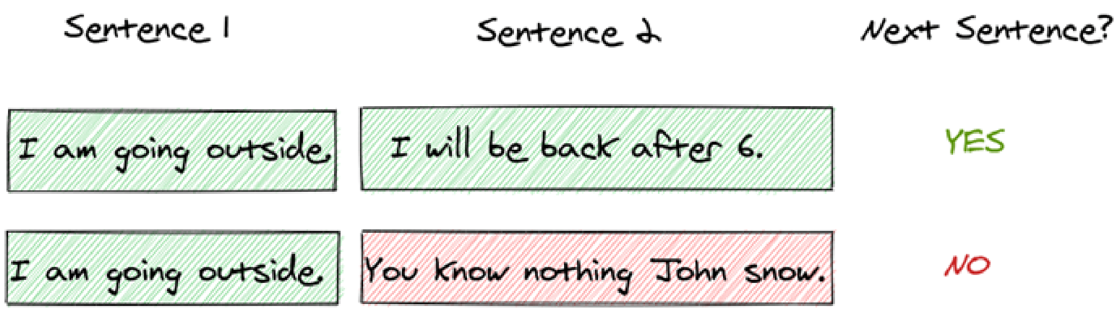
What is the main objective of the paper, what are they trying to solve?

Semantic similarity detection is a fundamental task in natural language understanding(NLP).
Adding topic information has been useful for previous feature-engineered semantic similarity models as well as neural models for other tasks.
There is currently no standard way of combining topics with pretrained contextual representations such as BERT. The paper investigates if topic models can further improve BERT's performance for semantic similarity detection. The objective is to show that combining topic information with BERT improves performance over various datasets.

Innovative part

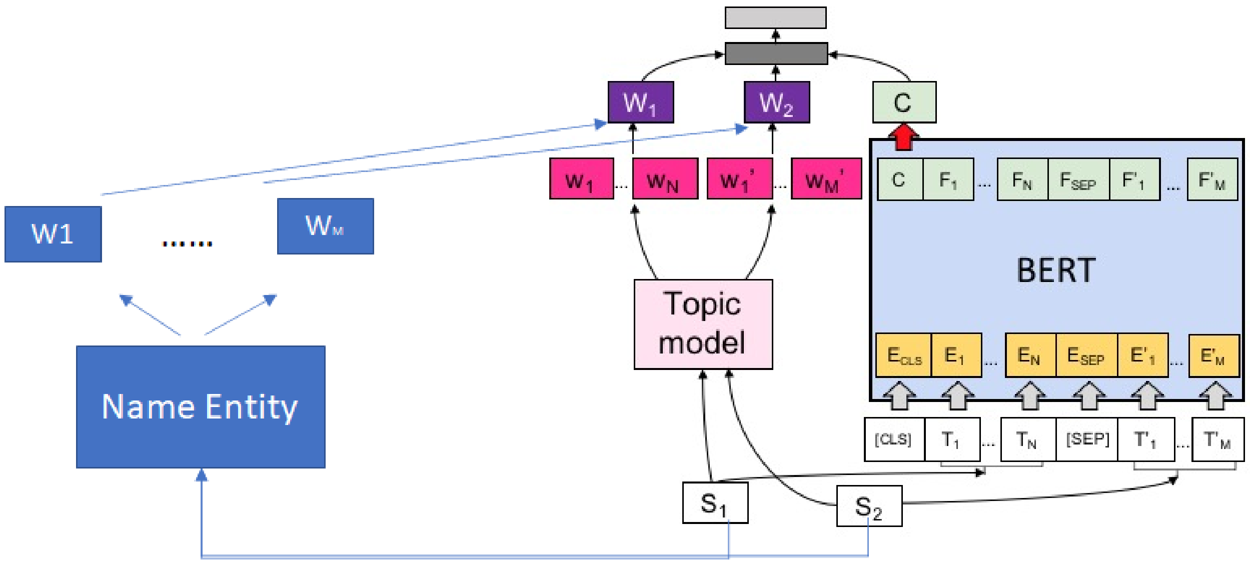
Clearly state the problem/issue/challenge that will address in your final project

- `Next Sentence Prediction`
  Using the tBERT model in order to check if it does a better job on a different task.



Sentence 1	Sentence 2	Next Sentence?
I am going outside,	I will be back after 6.	YES
I am going outside,	You know nothing John snow.	NO

- `Another research direction is to investigate if introducing more sophisticated topic models, such as named entity promoting topic models (Krasnashchok and Jouili, 2018) into the proposed framework can further improve results.` (As the article mentioned) So we will try to add a Name Entity Extraction as a parallel input for the LDA.
  For example:



Data sources - what are the data set(s) that you will use

We select two out of three popular benchmark datasets featuring in the article, with different sizes (small vs. large), type (QA vs. paragraph) and sentence lengths (short vs. long).

- **MSRP**
  Link Size: 5K
  The Microsoft Research Paraphrase dataset (MSRP) contains pairs of sentences from news websites with binary labels for paraphrase detection (Dolan and Brockett, 2005).
- **Quora**
  link Size: 404K
  The Quora duplicate questions dataset contains more than 400k question pairs with binary labels and is by far the largest of the datasets.

Evaluation - how are you going to evaluate performance

All of the above datasets provide two short text.
We frame the task as predicting the semantic similarity between two sentences in a **binary classification task**.

Metrics we'll use would be:

- Accuracy
- Precision
- Recall
- F1

Timeline and milestones

State at least 2 milestones (goals and dates) for the Anchor paper part

- 10/08: Datasets - Download, pre-procces, Data Loaders, etc.
- 17/08: Code: run tBERT's github code on Colab.

State at least 2 milestones (goals and dates) for the Innovative part

- 24/08: Run t-bert for a different task - `Next Sentence Prediction`.
- 31/08: Change t-bert architecture - add aditional `Name Entitiy Extraction` head.

Thanks for reading :)