# GraphNLI

ELAD LAKS (ID. 3115440084), OR KATZ (ID. 302269204)

Submitted as final project report
NLP course, IDC, 2022

## 1 Introduction

Talk-backs and debates of users on public discussion on forums can escalate into hate or misinformative posts.(1) To be able to tackle or to understand those posts the "GraphNLI" paper tries to classify whether reply is supporting or attacking the post it is replying to. The problem is that when trying to identify the polarity between post and reply, the reply to a post may be based on external context beyond the post. In order to explain the problem the paper define the "local context" as the context between post and it's reply, and "global context" as the context to all the other replies in and post connected to the local context. The paper suggest method to overcome the problem using the other replies and the replies to the reply from the 'root post'. They suppose it will give lot more information to decide whether the reply is supporting or attacking the post it is replying to. They named this method "GraphNLI". They compared several NLP models and methods on data-set from Kialo - a debate web, where every time user uploads a reply to post he first label his reply with attack/support flag. All of the methods based on GraphNLI was better then the others, and the best one was "GraphNLI: Root-seeking Graph Walk + Weighted Avg". It was inspired based on S-BERT using RoBERTa, Pooling, aggregation between the results making final embedding vector which is fed into Softmax-classifier in the end. The specialty in this method is the way the graph walks considering all the tree of replies and not only the reply and its post(considering the global context).

### 1.1 Related Works

#### 1.1.1 A natural language bipolar argumentation approach to support users in online debate interactions

link to the paper
One of the first papers that tried to predict support or attack context between post and reply published at 2013.(2) Obviously they were using old algorithms and tools. The paper proposing and evaluating the use of NLP techniques to identify the post and reply and their relations. In particular, they adopted textual entailment (TE) approach, a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. TE is then coupled together with an abstract bipolar argumentation system which allows to identify the arguments that are accepted in the considered online debate.
the paper archived accuracy of 69%.

#### 1.1.2 Identifying attack and support argumentative relations using deep learning

link to the paper
In this paper they propose a deep learning architecture to capture argumentative relations of attack and support from one piece of text to another, of the kind that naturally occur in a debate.(3)
The architecture uses two (unidirectional or bidirectional) Long ShortTerm Memory networks and trained word embeddings, and allowed to considerably improve upon existing techniques that use syntactic features and supervised classifiers for the same form of (relation-based) argument mining. The method was a deep learning architecture for Relation-based AM based on Long-Short Term Memory, within the architecture, each input text is fed, as a trained 100-dimensional GloVe embedding, into a (unidirectional or bidirectional) LSTM which produces a vector representation of the text independently of the other text being analysed. The two vectors are then merged (using element-wise sum or concatenation) and the resulting vector is fed to a softmax classifier which predicts whether the pair of input texts belongs to the attack, support or neither relations. the architecture achieved 89.53% (better from our paper).
As our paper mention, the data and the code of this paper is not available' so we couldn't compare between the models.

#### 1.1.3 GraphNLI

the unique of this paper is that in all of the above-mentioned approaches, the inputs to the model are the texts of the replying argument and the argument being replied to. Arguably, this is the least amount of information one must input into the model to predict the polarity of the reply. What has not yet been considered is whether it is helpful to input more information - the global context.

# 2 Solution

link to the paper git

## 2.1 The data-set

The data-set contains 324373 posts and replies, each one of them is a part of "conversetion tree". we split the data into train(80%) and test (20%). each conversation tree represent the global context, of local context between 2 sentences in the tree.

## 2.2 Graphnli method

#### 2.2.1 Distilroberta-base

The GraphNLI paper is mainly based on pre-trained model called distilroberta based on bert.(4) this is a distilled version of the RoBERTa-base model. On average DistilRoBERTa is twice as fast as Roberta-base. Roberta-base model created by changing the training of the original BERT, including new and larger data-set, far more iterations and much larger batches. It was pretrained with the Masked language modeling (MLM) objective and it without labeled data. This way, the model learns an inner representation of the English language that can then be used to extract features useful for downstream task


Figure 1: convesation

which makes it modular and robust. It trained on 160GB of text, 500K iterations, 8K batch size and 50K sub-words units.

#### 2.2.2 Garph walk architecture

In order to catch the global context without loosing the most important context(the local context) they were using graph walk method called - Weighted Root-seeking GraphWalk. Means to walk starting from a given node up towards the root of the discussion tree. They limited the walk to 5 steps, and weighted the nodes as follows constant(k = 0.75) power L = the number of steps from the given node $wi = k^{Li}$. At the end of a graph walk for each node, we obtain at most L=5 arguments which are ancestor of a given node. These sets of arguments are the input of the GraphNLI model.

#### 2.2.3 Model overview

Each of the L=5 arguments is input into the distilroberta-base model to get their corresponding embeddings,then, a mean-pooling operation is applied to derive a fixed-sized sentence embedding for each argument. They separated the given node(the point of interest) embaddings and the aggregated embeddings of its maximum L ancestors. The aggregation is with weighted average, according to the weight calculation explained in the previous paragraph. Then the embedding vectors concatenated with $|u - v|$ to get the final embedded vector which is fed into Softmax classifier for the prediction task.

### 2.3 design

We constructed the code on colab notebook. We stored the datasets on the drive, and load them from it to the notebook.

#### 2.3.1 Code mistakes in the original code

There was a few code mistakes in the original code, as follows:

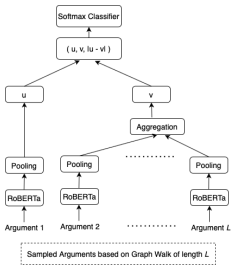


Figure 2: garph walk architecture

1. In gen_dataset_graph_walks - the construction of the dataframe from the graphwalk used maximum 4 sentences from each conversation. in order to walk with walklen= 5 they needed 6 sentences.

2. In SoftMaxLoss - they used mean aggregate instead of weighted_aggregate.

There was great effort to find code mistakes in the published version of the code – it was not using the right function and arguments, the graphwalk was to short and the code was not working.
After fixing all the errors, start to train distil-roberta model, as the paper asked. 4 epochs took 20 hours on gpu of colab pro machine (Tesla P100-PCIE-16GB) .

## 2.4 link to our notebook/git

github
colab

## 2.5 Paper results

the original paper used learning rate $2 * 10^{-5}$, warm-up over 10% of the data. batch size of 16, and Adam optimizer.

### 2.5.1 Results of the paper

The paper best result came from the combination of root seeking graph walk, with weighted avg. they have got 82.87% accuracy.

**Table 1: Accuracy scores of different models trained on Kialo dataset for polarity prediction, discussed in Section 5.4.**

| Model | Accuracy (%) |
|---|---|
| Bag-of-Words + Logistic Regression | 67.00 |
| Prompt Embeddings + Logistic Regression | 61.20 |
| Sentence-BERT with classifier layer | 79.86 |
| BERT Embeddings: Root-seeking Graph Walk + MLP | 70.27 |
| GraphNLI: Root-seeking Graph Walk + Sum | 80.70 |
| GraphNLI: Root-seeking Graph Walk + Avg. | 81.96 |
| GraphNLI: Root-seeking Graph Walk + Weighted Avg. | **82.87** |
| GraphNLI: Biased Root-seeking Random Walk + Sum | 79.95 |
| GraphNLI: Biased Root-seeking Random Walk + Avg. | 80.44 |

Figure 3: paper results

#### 2.5.2 our results

After 20 hours of training with the same hyper-parameters of the anchor paper except batch = 12 (because of memory limitations) we achieved 82.41% accuracy. This is a reasonable gap.

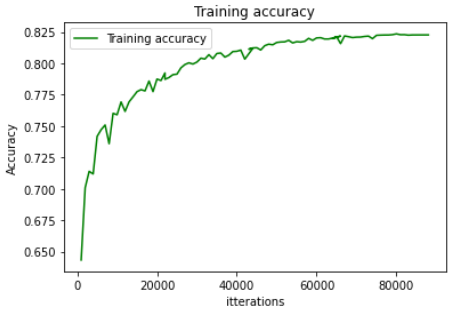


Figure 4: accuracy

# 3 Experimental results

## 3.1 Preprocess the data

We tried to preprocess the data better then the original paper. We removed stopwords and lower-cased all the data(because roberta-base is case sensitive).

## 3.2 Change the model

We used few other models in order to get better results. we used "sentence-transformers/all-mpnet-base-v2"(5), "cross-encoder/nli-deberta-v3-base"(6),"sentence-transformers/all-MiniLM-L6-v2". All of them was either really slow and we couldn't afford enough training time or got lower accuracy.

## 3.3 Change the aggregation method

In the original paper, as explained in the model overview(2.2.3) they used u,v, $|u - v|$. We wanted to check the influence of cosine similarity in the aggregation part but it hurt the accuracy.

## 3.4 Use the method on different dataset

We disided to try to use the GraphNLI method on different task. In the last part of the papr, they talked about future possible usage to the method, one of them was hate speech recognition. The main problem was to find dataset that fit to this mission. After grate effort of searching the web for the right dataset - with tree structure of post and reply, labeled to hate speech. We found this paper- "A Benchmark Dataset for Learning to Intervene in Online Hate Speech" (7) link to the paper which made auto reply to offensive language.
After fitting the dataset into our needs, we tried our method to hate speech prediction on the data of the paper, and got 91% accuracy.

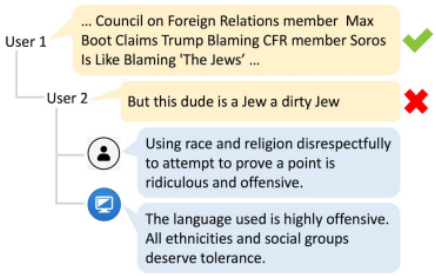


Figure 5: offensive language

# 4 Discussion

Classification of supporting or attacking relationship between post and reply found to be hard mission. The main reason is there is global context between the post and reply that does not take into consideration, even context that we can't find in the post tree such facts and knowledge. Our method tried to overcome part of this problem, and succeeded partly.we saw there are more possible usage to this method as hate speech recognition, and maybe even fake news classifier.
We belief the best way to overcome this problem of understanding the relationsheep between post and reply is to practice and learn large datasets, with updated facts to teach the model the largest global context he can.

# 5 Code

colab notebook

# References


[1] Agarwal, Vibhor, Sagar Joglekar, Anthony P Young, and Nishanth Sastry: *Graphnli: A graph-based natural language inference model for polarity prediction in online debates*. In *Proceedings of the ACM Web Conference 2022*, pages 2729–2737, 2022.

[2] Cabrio, Elena and Serena Villata: *A natural language bipolar argumentation approach to support users in online debate interactions†*. Argument & Computation, 4(3):209–230, 2013. https://doi.org/10.1080/19462166.2013.862303.

[3] Cocarascu, Oana and Francesca Toni: *Identifying attack and support argumentative relations using deep learning*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. https://aclanthology.org/D17-1144.

[4] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf: *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*. ArXiv, abs/1910.01108, 2019.

[5] Reimers, Nils and Iryna Gurevych: *Sentence-bert: Sentence embeddings using siamese bert-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019. http://arxiv.org/abs/1908.10084.

[6] He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen: *Deberta: Decoding-enhanced bert with disentangled attention*. In *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=XPZIaotutsD.

[7] Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang: *A benchmark dataset for learning to intervene in online hate speech*, 2019. https://arxiv.org/abs/1909.04251.

[8] Chi Sun, Xipeng Qiu, Yige Xu Xuanjing Huang: *How to fine-tune bert for text classification?* 2020. https://arxiv.org/pdf/1905.05583.pdf.