HOW TO DISAGREE
REPORT

Marco Lassandro ID. 945907

1 INTRODUCTION

The web is turning writing into conversations, there are many discussions about any topic on socials, forum, blogs and on many other web sources. In almost all the conversations there is always someone that disagree on something that someone else stated about a certain topic; is an expected behavior because agreeing on something motivate people less than disagreeing and when you agree there is less to say.

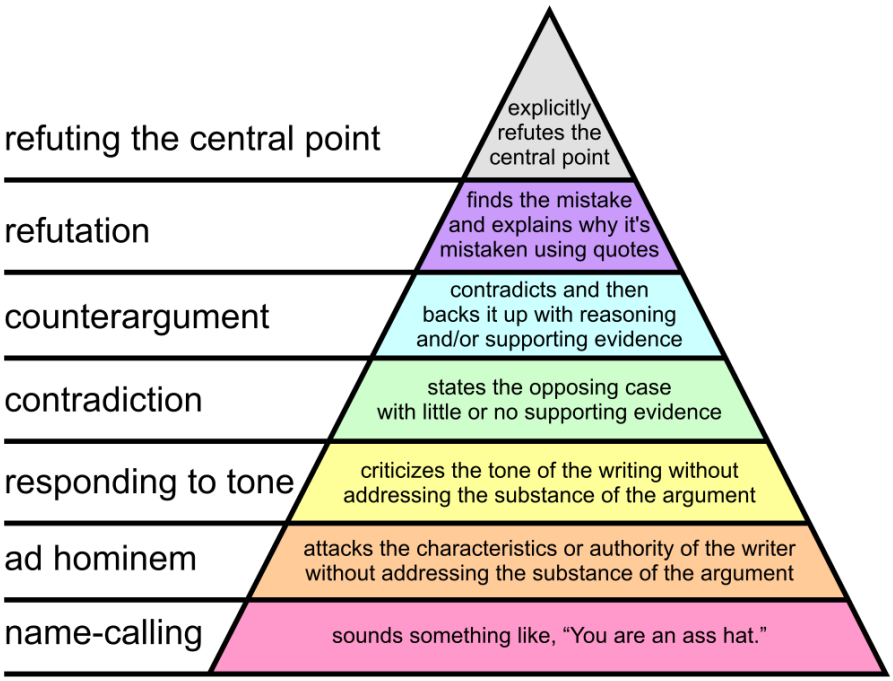
If we're all going to be disagreeing more, we should be careful to do it well; for these reasons could be useful to have a machine learning model that can detect how accurate is a disagreement with respect to something said before. The project aims to analyse paired statements that are related to a common topic on different debates, where the second statement is in disagreement with respect to the first of the pair. The paired statements are then classified according to a derivation of the disagreement hierarchy defined by Paul Graham.

In order to do so, a dataset has been built by using a mixture of different dataset, resulting in a set of 950 examples; then for the classification task, two similar neural network models have been built and trained; the two architectures have been extracted from this [1]. To prepare properly the examples, in order to be passed through the networks, different pre-processing techniques have been applied over the texts. Two metrics have been used to evaluate the error over the classifications performed by the models, that are: the accuracy and the AUROC.

2 EXPERIMENTAL SETUP

2.1 THE PAUL GRAHAM DISAGREEMENT HIERARCHY

Paul Graham is a programmer, writer and investor that defined a hierarchy of disagreement, here below we can see a visualisation of such hierarchy:



(a)

Figure 1: Original structure of the disagreement hierarchy defined by Paul Graham

As we can see here the disagreement hierarchy has a pyramidal structure, this means that the lowest levels are the most common disagreements that we can find in the web and as we get to the top levels, we get more rare and interesting disagreements. In the first three levels there are all the disagreements that are not related to the content of what has been stated before but rather is an attack on the person or on the way a thing has been stated; in the upper levels, instead, there are disagreements related to the content of what has been stated.

2.2 THE DATASET

The dataset, has said in the introduction, is a collection of 950 examples labeled according to a variant of the Paul Graham hierarchy explained before. It has been stored as a csv file, here below an explanation of the fields:

- statement_1: the first statement refers to a given topic;
- statement_2: the second statement is the disagreement given to the first statement;
- label: the label is an integer, in a range from 1 to 4, that represents a category of the disagreement;

2.2.1 VARIANT OF THE PAUL GRAHAM HIERARCHY

In order to simplify the labeling of the examples, it has been decided to merge the first two levels of the Paul Graham hierarchy and the last two levels, obtaining in this way four categories of disagreements:

- 1st category: all the disagreements that are not related to the content of what has been stated before, but rather on the person;
- 2nd category: all the disagreements that state just the opposite of what has been stated before;
- 3rd category: all the disagreements that state the opposite and make a reasoning by adding some evidence on their side.
- 4th category: all the disagreements that make refutations of what has been stated before, generally the refutation is done by quoting some passages of the previous statement and explaining why it's mistaken.

In this project the "Responding to Tone" level has been not considered.

2.2.2 CONSTRUCTION OF THE DATASET

Three different dataset has been used in order to build the final dataset, this has been done in order to simplify the construction of the various examples for each category. Here below an explanation of the various dataset:

- Agreement by Create Debaters (ABCD): The ABCD corpus was built from the Create Debate website where users can start a debate by asking a question. Although the website can support open ended as well as multiple sided debates.
- Hate Speech and Offensive Language Dataset: This dataset has been retrieved from a kaggle competition called "Detecting Insults in Social Commentary", it is a competition where the challenge is to detect when a comment from a conversation would be considered insulting to another

participant in the conversation. The data consists of a label column followed by two attribute fields. The label is either 0 meaning a neutral comment, or 1 meaning an insulting comment (neutral can be considered as not belonging to the insult class. The first attribute is the time at which the comment was made, the second attribute is the unicode-escaped text of the content, surrounded by double-quotes. The content is mostly english language comments, with some occasional formatting. For the project just the examples with labels equal to 1 has been considered.

- Contradiction dataset: This dataset has been retrieved from a kaggle competition called "Contradictory, My Dear Watson", it consists in creating an NLI model that assigns labels of 0, 1, or 2 (corresponding to entailment, neutral and contradiction) to pairs of premises and hypotheses. In this case only the examples , with label equal to 2 have been considered, moreover all the examples, that have a language different from the english, have been filtered out;

For each category explained above 235 examples have been retrieved by using a mixture of such dataset, in particular:

- 1st category: in this case the first statements of the pairs have been retrieved from the ACBD dataset, meanwhile the second statement of the pairs have been retrieved from the "Hate Speech and Offensive Language Dataset". This has been done in order to simulate the name-calling/ad-hominem disagreement level.

- 2nd category: for the second category the pairs have been retrieved from the "Contradiction dataset";

- 3rd and 4th category: for the last two categories just the ACBD dataset has been used;

3 EXPERIMENTAL SETUP

In this section will be presented the pre-processing steps and the architectures of the machine learning models used for the experiments on the given classification task.

These are the text preprocessing / cleaning steps used:

- Lower casing;

- Removal of punctuations;

- Removal of stopwords;

- Removal of HTLM tags;

- Expansions of verbal contractions;

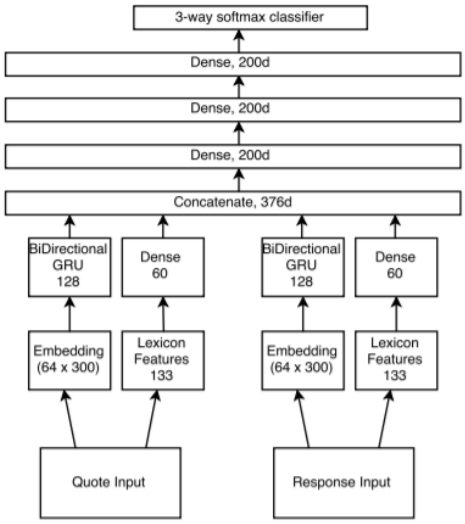
After these steps a vectorization of the statements have been performed in order to map text features to integer sequences, the sequences length have been fixed to 200, this has been decided looking to an histogram of the lengths. In case of statements with length lower than 200, a zero-padding has been added to the sequences.

3.1 WORD EMBEDDING: GloVe

GloVe is essentially a log-bilinear model with a weighted least-squares objective. The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. It has been exploited a pre-trained model of GloVe to embed the words of the statements, in particular the one with 300 dimensions trained on Common Crawl with 840 billion tokens.

3.2 MACHINE LEARNING MODELS

The architectures for the two models used in the experiments have been inspired by this [1]. A visualisation of their neural network architecture can be seen here:



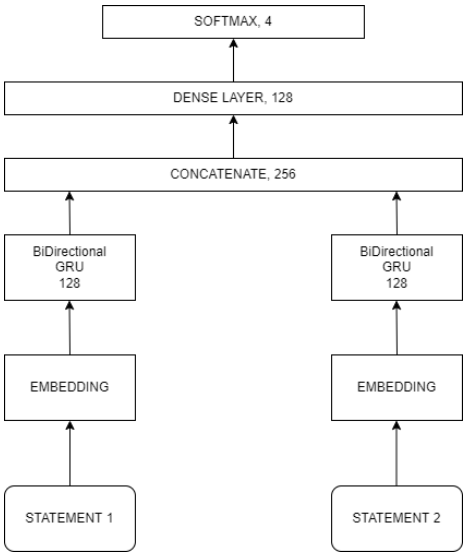
(a)

Figure 2: Neural Network architecture of the [1]

This model has been used to detect (dis)agreement by performing a 3-way classification (agreement/disagreement/none) between the Q-R pairs on several

existing annotated dataset. For each Q-R pair they extract two sets of features. First, GloVe word embeddings are fed to Gated Recurrent Unite to create a sentence embedding. Second, from each text a lexical feature vector is extracted; they have used affect, sentiment, emotion, opinion lexicons for feature extraction because in many of the online discussions forums people tend to argue with emotion and opinion about a particular topic to convey their stance or belief. Both these sentence embeddings from Q-R pairs are concatenated and then fed into fully connected layers to do 3 way classification.

The above architecture has been simplified and adapted for our task, here below a view:

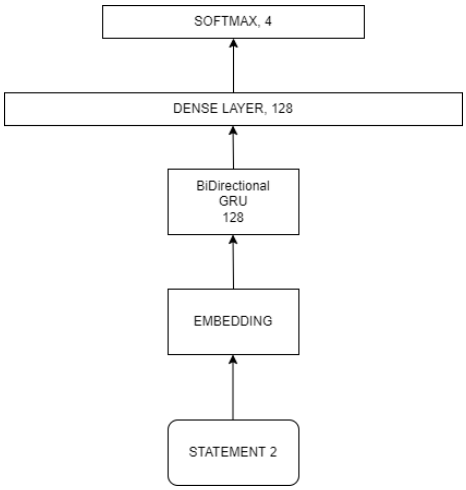


(a)

Figure 3: Architecture used for the disagreement classification task

The lexicon feature has been not used and the number of fully connected layers has been reduced to 1, this has been decided for the fact that we have a pretty small dataset; also the number of neurons for the fully connected layer has been reduced to 128. Eventually the number of neurons for the softmax layer has been passed to 4, since we have a 4 way classification.

The second architecture model can be seen in this picture:



(a)

Figure 4: Architecture used for the disagreement classification task

In this case there is no presence of a second GRU branch, so as input we have just the second statement, that is the one related to the disagreement; this has been done because there was a need to understand how much impact the presence of the first statement had over the results.

3.3 SYSTEM PARAMETERS

For both the models, the fully connected layers have a ReLU activation; the network is optimized with Nadam optimizer with learning rate of 0.001 and the loss function used was the "categorical_crossentropy".

3.4 RESULTS

In order to evaluate the performances of the models over the created dataset, a stratified K-fold cross validation has been performed in order to obtain the same percentage examples per category in every fold; the k parameter has been fixed to 10, so we have 95 examples in the test fold. For each fold 20 epochs had been run over the optimization phase of the models.

The metrics used to evaluate the performances had been the accuracy and the AUROC metric.

For each model, here we can see the results:

Model	Accuracy	AUROC
Single GRU	0.85	0.96
Dual GRU	0.95	0.99

Table 1: Best results obtained on the regression task.

Given the best model trained over the K-fold validation we can see below a visualisation of the metrics' values at each epoch on the optimization phase:

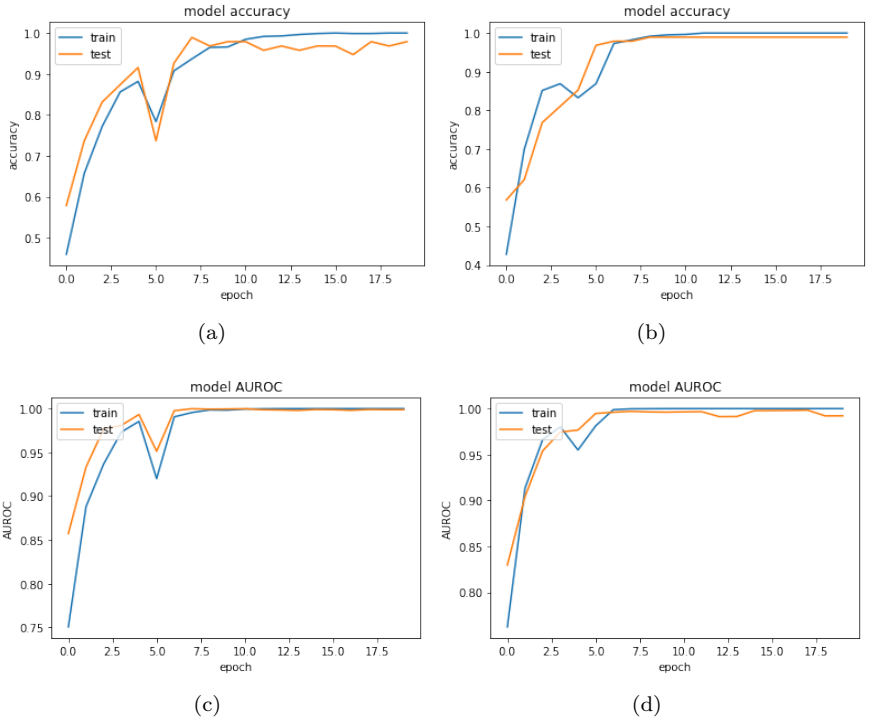


Figure 5: (a) Accuracy history single GRU model (b) Accuracy history dual GRU model (c) AUROC single GRU model (d) AUROC dual GRU model

3.5 CONCLUSIONS AND FUTURE WORKS

As we can see the results seems to be promising, the architecture proposed by [1] with the usage of the two branches is very effective for our task. In the future could be useful to make an augmentation of the examples for each category; moreover could be interesting apply this type of solution as a sort of filter, for example on the comment section of social websites, letting the users decide the level of disagreement they want to see on any discussion of the given social.

References

[1] Sushant Hiray and Venkatesh Duppada. Agree to disagree: Improving disagreement detection with dual grus. volume abs/1708.05582, 2017.