Automatic detection of persuation attempts on social networks

Ruben Teimas

m47753@alunos.uevora.pt

Departamento de Informática
Escola de Ciências e Tecnologia

September 9, 2023

UNIVERSIDADE DE ÉVORA

Introduction

State of the art

Problem study

Proposed system

Results

Future work

Introduction

- In 2021, a US citizen spent on average 1300 hours on Social Networks.
- Social Network's nature allow unmoderated content to be created and shared.
- Content on social media can influence public opinions.
- SemEval challenge as a starting point.
- Create an intelligent open-source system for content moderation on social networks.

Introduction

State of the art

Problem study

Proposed system

Results

Future work

State of the art

- Investigated the problem at hands at different levels.
- Researched about techniques to help us achieve our goals.
- Investigated multi-label classification problems.
- Dug into other participants submissions.

Introduction

State of the art

Problem study

Proposed system

Results

Future work

Data analysis

- Data collected from Facebook groups regarding politics, covid and gender equality.
- Data consisted of 950 total entries:
  - **Train**: 687.
  - **Dev**: 63.
  - **Test**: 200.

Data analysis

Persuasion Techniques	Sub-task 1	Sub-task 3
Loaded Language	489	761
Name Calling/Labeling	300	347
Smears	264	602
Doubt	84	111
Exaggeration/Minimisation	78	100
Slogans	66	70
⋮	⋮	⋮
Straw Man	24	40
Appeal to Authority	22	35
Reductio ad Hitlerum	13	23
Obfuscation, Intentional Vagueness, Confusion	5	7
Presenting Irrelevant Data (Red Herring)	5	7
Bandwagon	5	5
Total	**1642**	**2488**

Table: Persuasion technique's statistics.

Data analysis

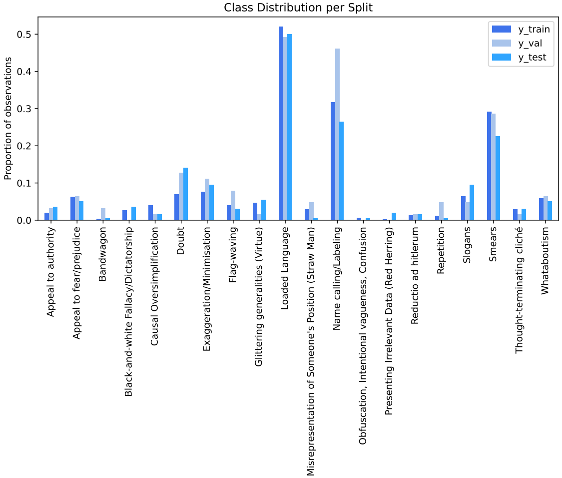

Figure: Class distribution per split on *sub-task 1*.

Text pre-processing pipeline:

- Case converting.
- Pos-Tagging.
- Tokenization.
- Stop-Word removal.
- Lemmatization.

	Nº of Words	Nº of distinct words
Unprocessed text	16840	6427
Pre-processed text	9483	3092

Table: Corpus dimension before and after pre-processing

ML Models

UNIVERSIDADE DE ÉVORA

- Modeled the problem with problem transformation techniques such as Binary Relevance and Label Powerset.
- Used Tf-Idf and Word2vec for feature extraction.
- A combination of Train and Dev set were used to train the models.
- The models were evaluated using K-Cross validation, with $K = 5$.

ML Models

- The results were not very good using either TF-IDF and Word2vec.
- The best results came from using Word2vec with Naive Bayes and Binary Relevance, with 0.36 of Micro F1-score and 0.19 Macro F1-score.
- We decided to use more sophisticated approaches.

Introduction

State of the art

Problem study

Proposed system

Results

Future work

Proposed system

- DistilBERT reduces BERT's by 40% while retaining 97% of its functionality.
- Convolutional layers can help recognize patterns in the sentences.
- Convolution is easy to compute on a GPU due to memory structure.
- Dropout layers help to prevent overfitting.

Architecture

UNIVERSIDADE DE ÉVORA

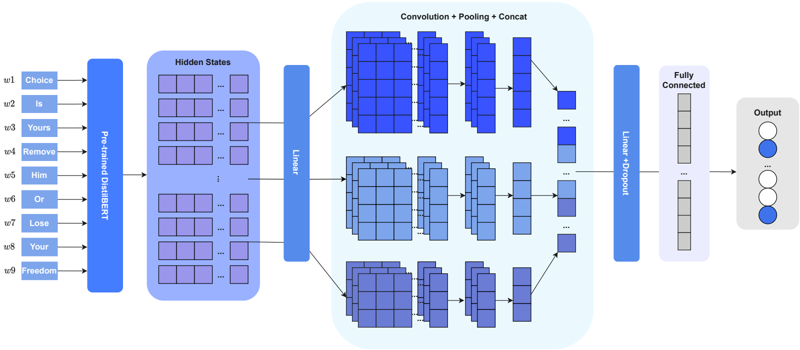

Figure: Final model architecture.

Experiment's methodology

UNIVERSIDADE DE ÉVORA

- Train models on Train and Dev sets.
- Validation is performed using stratified k-cross validation, with $K = 5$.
- Model Tweaking:
    - Loss function.
    - Text pre-processing.
    - DistilBERT's fine-tuning.
    - Hyperparameter search.

Introduction

State of the art

Problem study

Proposed system

Results

Future work

Baseline system

- AdamW optimizer.
- Binary Cross Entropy as loss function.
- Micro f1-score of **0.516** and **0.116** macro f1-score.
- Inability to predic 12 out of 20 classes.

Hyperparameter	**Value**
Learning rate	3e-5
Epochs	10
Batch size	8
Dropout rate	0.2
Filters	128
Kernel dims.	[3,4,5]
Hidden dim.	768

Table: Baseline system hyperparameters.

Text pre-processing

- Using the parameters from baseline system.
- Applied the pre-processing text pipeline previously created.
- Worst results for every class.
- DistilBERT can benefit of text being in its natural form.

Loss functions

- Binary Cross Entropy (BCE) computes the same weights for all class samples.
- Excessive focus on learning the most represented classes with a large number of training sample.
- Focal Loss (FL) introduces a modulating factor $(1 - p_t)^\gamma$ to BCE.
- As p_t gets closer to 1 the factor goes to 0 and the loss for well-classified examples is down-weighted.

UNIVERSIDADE DE ÉVORA

- Setting a high γ, to sufficiently down-weight the contribution from easy negatives, may eliminate the gradients from the rare positive samples.
- Asymmetric Loss (ASL) overcomes this problem by decoupling the focusing levels of the positive and negative samples.

	Micro F1-score	**Macro F1-score**
Binary Cross Entropy	0.516	0.116
Focal Loss	0.523	0.162
Asymmetric Loss	0.525	0.238

Table: Micro and Macro F1-scores for different loss functions.

DistilBERT fine-tuning

Model	Micro F1-Score	Macro F1-Score	Parameters (Millions)
F0	0.525	0.238	68
F1	0.519	0.252	61
F2	0.515	0.244	54
F3	0.515	0.230	47
F4	0.505	0.241	40
F5	0.505	0.234	33
F6	0.503	0.177	26

Table: Macro and Micro F1-Score freezing DistilBERT's layers.

Hyperparameter Search

- We Used Tree Parzen Estimators (TPE) for hyperparameter search.
- TPE uses probabilistic models to guide the search.

Hyperparameter	**Type**	**Values**
Learning rate	Float	$[1e-6, 1e-4]$
Dropout rate	Float	$[0.05, 0.45]$
Number of filters	Int	$i \in [5, 8]$ for 2^i
Kernel dimensions	Choice	$[[1, 2], [3], [3, 4, 5]]$
Hidden layer dimension	Int	$i \in [8, 11]$ for 2^i

Table: Search space for hyperparameters.

Hyperparameter Search

UNIVERSIDADE DE ÉVORA

Model	Learning Rate	Dropout	Filters	Kernels	Hidden Layer Dim	Micro F1	Macro F1
H1	1.189e-04	0.175	128	[3,4,5]	256	0.528	0.201
H2	6.790e-05	0.119	128	[3]	512	0.539	0.272
H3	1.879e-05	0.159	64	[3]	1024	0.504	0.240
H4	6.926e-05	0.103	64	[1,2]	512	0.536	0.246
H5	3.353e-05	0.373	64	[3,4,5]	1024	0.547	0.223
H6	4.813e-05	0.063	128	[3]	1024	0.556	0.246
H7	3.993e-05	0.095	128	[3,4,5]	1024	0.551	0.233
H8	6.925e-05	0.107	64	[3,4,5]	256	0.499	0.261
H9	1.578e-05	0.305	128	[3,4,5]	256	0.519	0.231
H10	2.433e-05	0.144	128	[1,2]	1024	0.555	0.212

Table: Models trained using TPE as search algorithm.

Analyzing overfitting

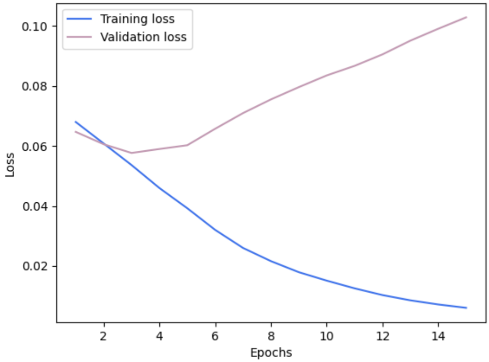

Figure: *H6* model loss evolution.

Analyzing overfitting

- Plotted the loss evolution using the different loss functions.
- Removed the convolutional layers from the model and increased the dropout rate to 20%.
- Froze all the distilBERT layers (except for the last one).
- Ended up training 3 models:
  1. H6 model for 4 epochs.
  2. H6 model for 15 epochs.
  3. H6 model with a learning rate ($4\mathrm{e}{-6}$) for 15 epochs.

Discussion

UNIVERSIDADE DE ÉVORA

	F1-Score			
Technique	**MinD**	**H6.1**	**H6.2**	**H6.3**
Appeal to Authority	0	0.333	0.545	0
Appeal to Fear/Prejudice	0.522	0.333	0.333	0.278
Bandwagon	0	0	0	0
Black-and-White Fallacy/Dictatorship	0.400	0	0	0
Causal Oversimplification	0.500	0.286	0.222	0
Doubt	0.400	0.387	0.340	0.378
Exaggeration/Minimisation	0.550	0.375	0.542	0.333
Flag-Waving	0.615	0.286	0.444	0.316
Glittering Generalities (Virtue)	0.286	0.190	0.222	0.174
Loaded Language	0.819	0.813	0.823	0.805

Table: Pt.1 Comparing the final models results by class.

Discussion

	F1-Score			
Technique	**MinD**	**H6.1**	**H6.2**	**H6.3**
Straw Man	0	0	0	0
Name Calling/Labeling	0.667	0.600	0.600	0.592
Obfuscation, Intentional Vagueness, Confusion	0	0	0	0
Presenting Irrelevant Data (Red Herring)	0	0	0	0
Reductio ad Hitlerum	0	0	0	0
Repetition	0	0	0	0
Slogans	0.154	0.303	0.250	0.242
Smears	0.511	0.468	0.486	0.467
Thought-Terminating Cliché	0	0	0	0
Whataboutism	0.375	0.190	0.333	0.292

Table: Pt.2 Comparing the final models results by class.

UNIVERSIDADE DE ÉVORA

Rank	Model	Micro F1-Score	Macro F1-Score
1	MinD	0.593	0.290
2	Volta	0.570	0.262
3	**H6.2**	0.551	0.257
5	AIMH	0.539	0.245
6	**H6.1**	0.526	0.228
7	DistilBERT	0.515	0.251
8	LeCun	0.512	0.227
11	**H6.3**	0.509	0.198
12	NLyticsFKIE	0.498	0.140
13	RoBERTa	0.497	0.240
16	YNUHPCC	0.493	0.263
19	NLPIITR	0.379	0.126

Table: SemEval's systems comparison.

Introduction

State of the art

Problem study

Proposed system

Results

Future work

Future work

UNIVERSIDADE DE ÉVORA

- Augment the dataset using other SemEval data.
- Build ensemble of classifiers.
- Explore multimodal approaches.

Thank you!

Thank you for your attention,

Ruben Teimas

GitHub github.com/TeimasTeimoso
LinkedIn linkedin.com/in/ruben-teimas