Offensive Language Detection in English and Greek

Yi-Chien Lin, Bridget Tyree, David Yi, Levon Haroutunian

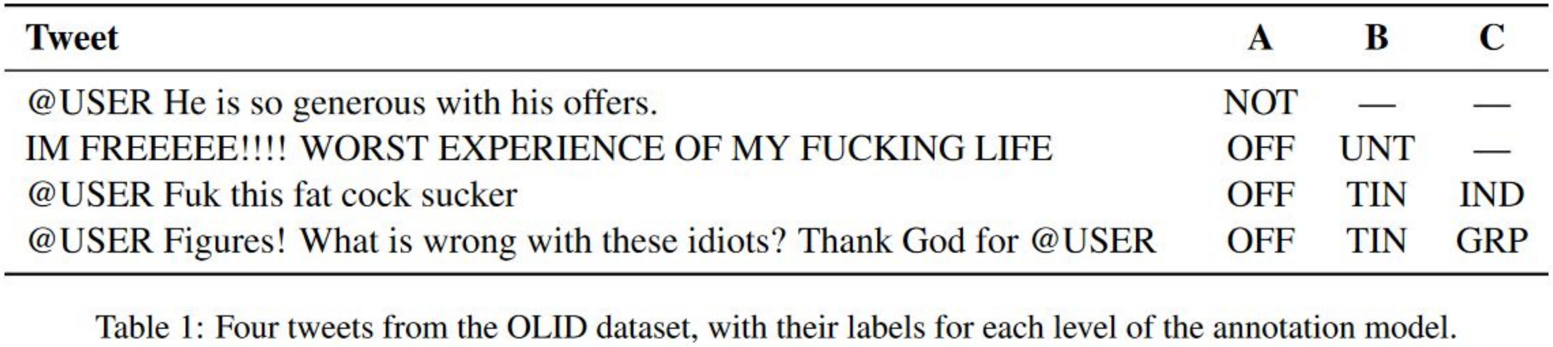
Primary task

Subtask A of OffensEval 2019

Binary classification: is an English tweet offensive (OFF) or not offensive (NOT)?

Data: OLID (Offensive Language Identification Dataset)

Total: 14,100 tweets – 9,460 NOT & 4,640 OFF tweets



Tweet	A	B	C
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Table 1: Four tweets from the OLID dataset, with their labels for each level of the annotation model.

From Zampieri et al., 2019

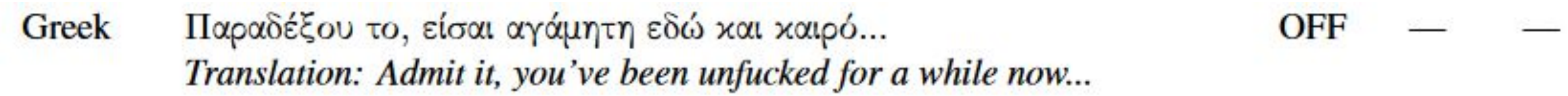
Adaptation task

Subtask A of OffensEval 2019 with Greek data from OffensEval 2020

Binary classification: is an Greek tweet offensive (OFF) or not offensive (NOT)?

Data: OGTD (Offensive Greek Tweet Dataset)

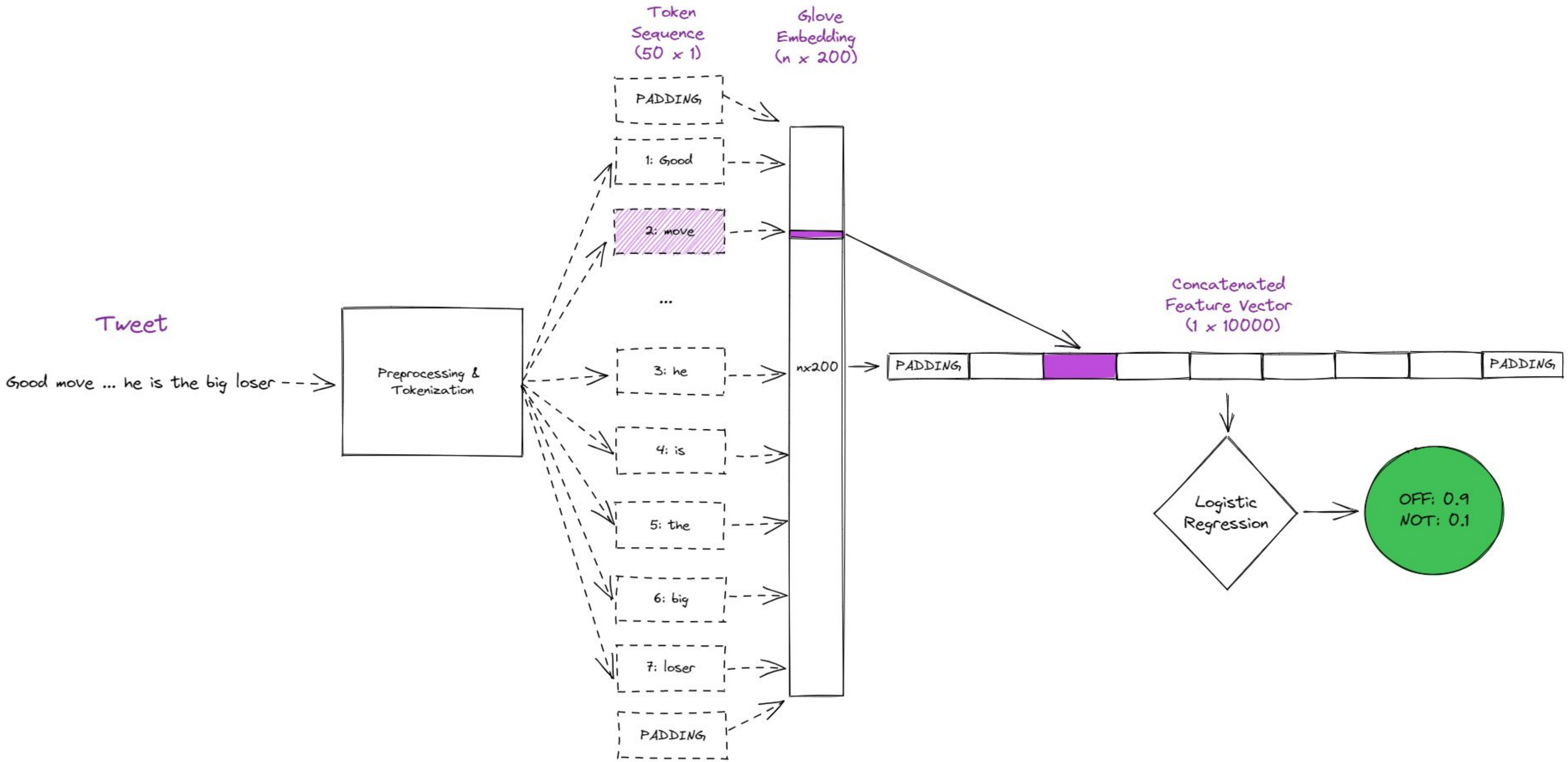
Total: 10,287 tweets – 7,376 NOT & 2,911 OFF tweets



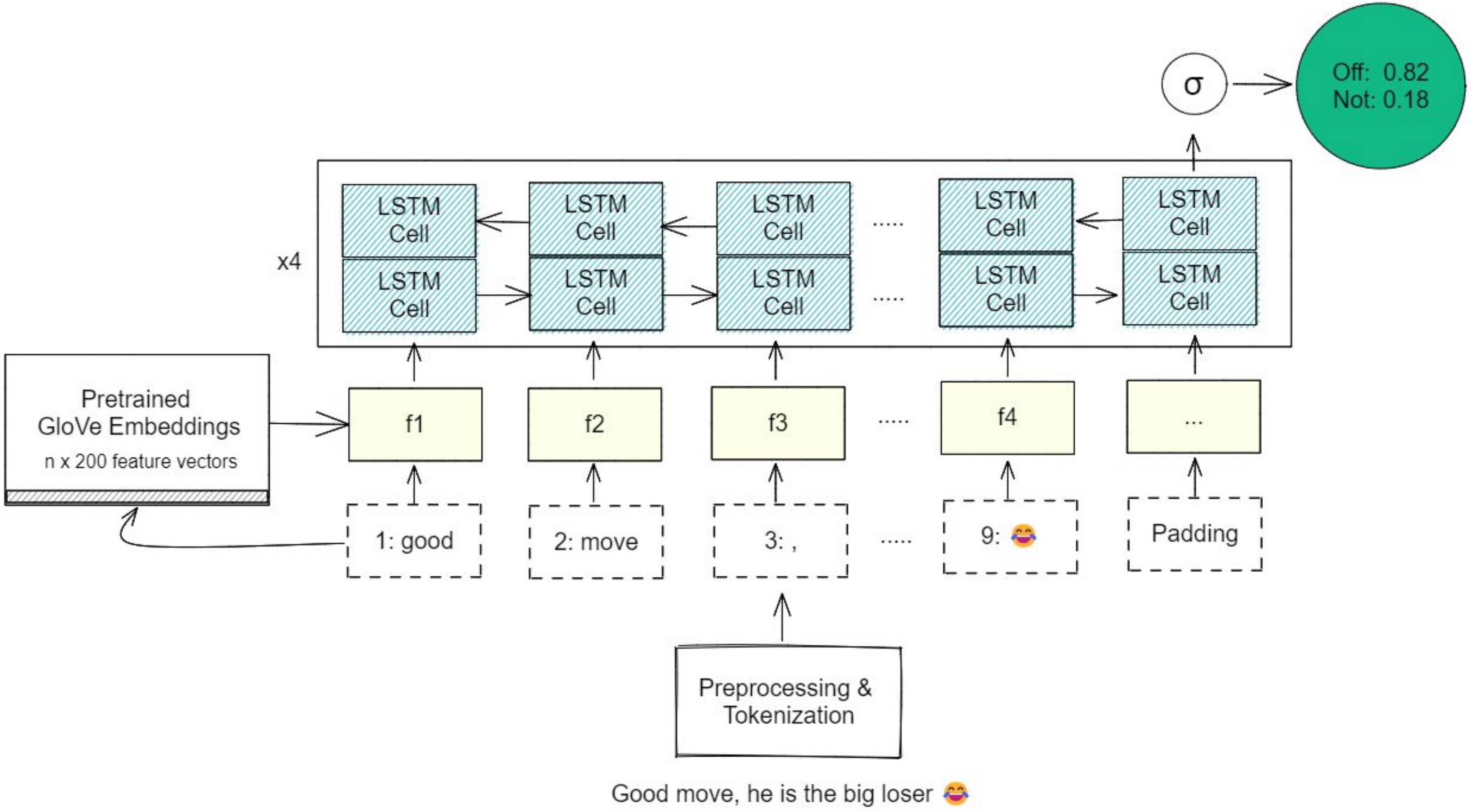
Greek	Παραδέξου το, είσαι αγάμητη εδώ και καιρό... *Translation: Admit it, you've been unfucked for a while now...*	OFF	—	—

From Zampieri et al., 2020

Overview of Previous Systems: D2



Overview of Previous Systems: D3



Current System (Core Approaches)

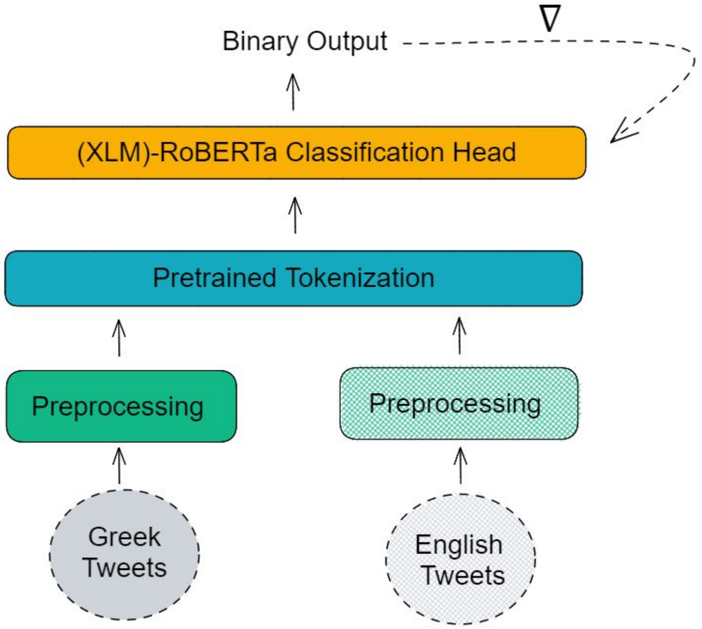
Previously: Logistic Regression + GloVe (D2) → BiLSTM + GloVe (D3)

Now:

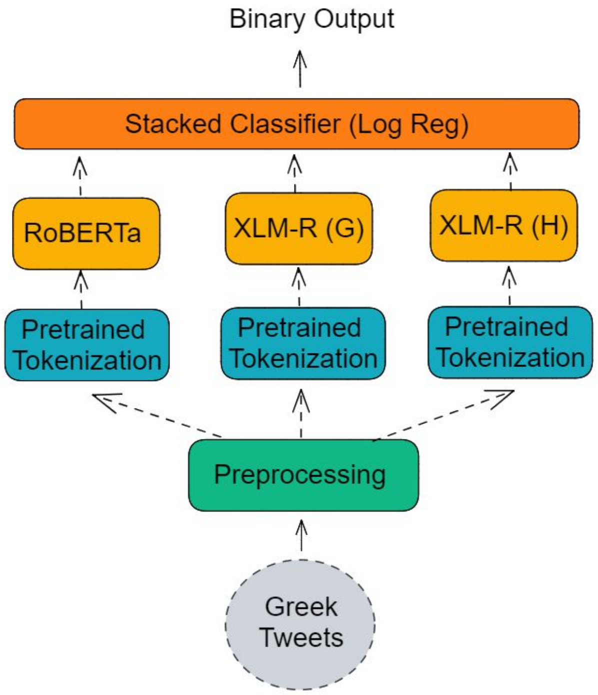
- Language-specific pre-processing for English and Greek
- BERT-style tokenizer and encoder (mBERT, RoBERTa, XLM-RoBERTa)
- Fine-tuned final classification layer (binary)
- Hybrid training: using data from both languages to train, then evaluating separately for each language
- Ensembling to combine predictions from multiple models

Current System (Architecture)

Hybrid Finetuning (Gr/En)



Ensemble Prediction (Gr)



Pre-processing (Primary Task): English

- Converting tweets into lowercase
- Splitting punctuations
- Removing apostrophes
- Removing hashtags

By the way, I don’t agree with your argument. #livid

↓

by the way , i dont agree with your argument . livid

Pre-processing (Adaptation Task): Greek

- Basic Preprocessing
  - Splitting punctuations
  - Removing apostrophes
  - Removing hashtags
- Removing Diacritics
  - ά → *α*
- Converting unicode data into ASCII characters
  - που → *pou*
- Lemmatization
  - spaCy
  - Converting words to their base form
  - Example (in English): walks/walked/walking → *walk*

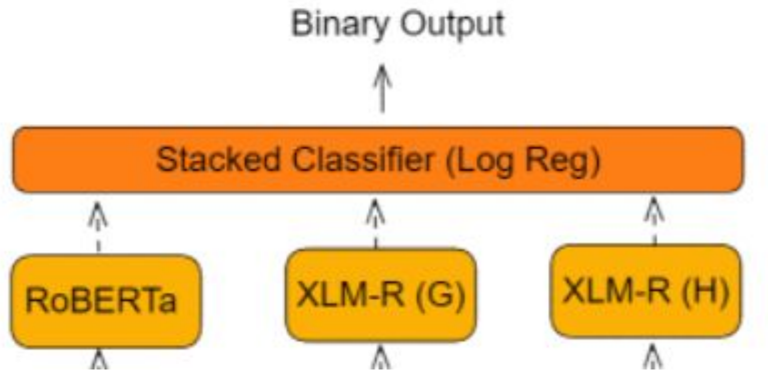
Additional methods	Validation F1 score	Test F1 Score
N/A	0.825	0.682
Remove diacritics	0.829	0.672
Convert ascii	0.720	0.671
Lemmatize	0.815	0.673
Lemmatize + Remove diacritics	0.822	0.658

Ensembling

Two relatively simple methods for ensembling

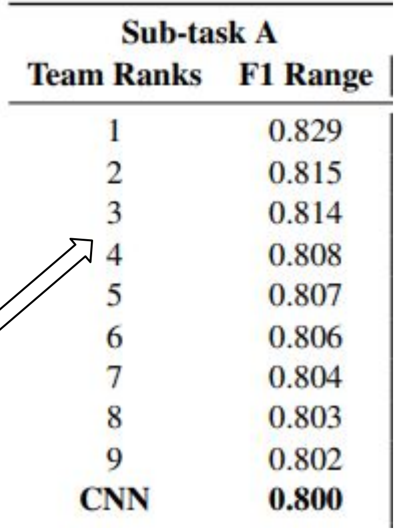
Averaging Ensemble: Average the probabilities of each individual model to get the final prediction

Stacked Classifier: Train a Logistic Regression "meta-learner" that's trained on the prediction probabilities of each individual model.



Primary Task (2019): Results

Approach	Dev Macro F1	Test Macro F1
All OFF	N/A	0.220
All NOT	N/A	0.420
D2 Model - Log Regression + GloVe	0.598	N/A
D3 Model - BiLSTM + GloVe	0.729	N/A
D4 Model - RoBERTa (en/gr)	0.790	0.787
D4 Model - XLM-R-Large (en)	0.782	**0.809**
D4 Averaging Ensemble	0.789	0.808
Best (BERT-base-uncased)	N/A	0.829



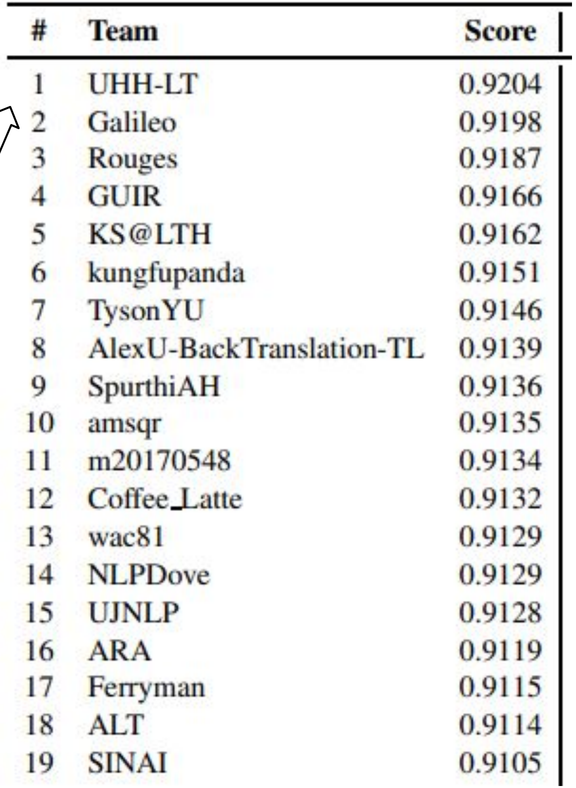
Sub-task A	
Team Ranks	F1 Range
1	0.829
2	0.815
3	0.814
4	0.808
5	0.807
6	0.806
7	0.804
8	0.803
9	0.802
CNN	**0.800**

Primary Task (2020): Results

OLID 2020 English Dataset with additional semi-supervised test examples:

train:	test:
NOT: 8840 (0.668)	**NOT**: 2807 (0.722)
OFF: 4400 (0.332)	**OFF**: 1080 (0.278)

Approach	Validation F1 Score	Test F1 Score
Stacked Ensemble: RoBERTa (en/gr) + XLM-R (en) + XLM-R-Large (en)	N/A	**0.9202**
RoBERTa (en/gr)	0.790	**0.9156**
Majority Baseline	N/A	**0.4193**



#	Team	Score
1	UHH-LT	0.9204
2	Galileo	0.9198
3	Rouges	0.9187
4	GUIR	0.9166
5	KS@LTH	0.9162
6	kungfupanda	0.9151
7	TysonYU	0.9146
8	AlexU-BackTranslation-TL	0.9139
9	SpurthiAH	0.9136
10	amsqr	0.9135
11	m20170548	0.9134
12	Coffee_Latte	0.9132
13	wac81	0.9129
14	NLPDove	0.9129
15	UJNLP	0.9128
16	ARA	0.9119
17	Ferryman	0.9115
18	ALT	0.9114
19	SINAI	0.9105

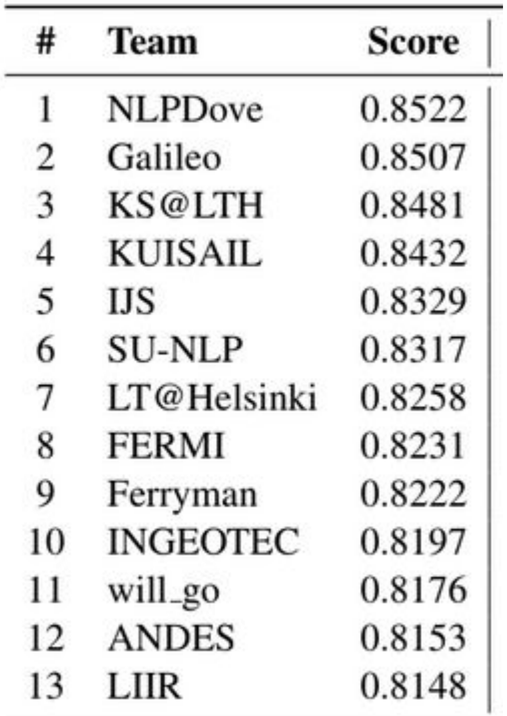
Adaptation Task: Results

Offensive Language Detection In Greek

train:
NOT: 6257 (0.716)
OFF: 2486 (0.284)

test:
NOT: 988 (0.689)
OFF: 446 (0.311)

Approach	Validation F1 Score	Test F1 Score
NLPDove (First Place Team)	N/A	0.8522
Greek Stack Classifier: XLM-R-Large (gr) + XLM-R-Large (en/gr)	N/A	**0.7015**
XLM-R-Large (gr)	0.825	**0.6815**
Majority Baseline	N/A	0.4202



#	Team	Score
1	NLPDove	0.8522
2	Galileo	0.8507
3	KS@LTH	0.8481
4	KUISAIL	0.8432
5	IJS	0.8329
6	SU-NLP	0.8317
7	LT@Helsinki	0.8258
8	FERMI	0.8231
9	Ferryman	0.8222
10	INGEOTEC	0.8197
11	will_go	0.8176
12	ANDES	0.8153
13	LIIR	0.8148

Challenges and Successes

Challenges:

- Tried to do too much for D3
- None of us speak Greek
- Gap between Greek dev and test F1
- Class Imbalance

Successes:

- Scored high enough to achieve rankings of
  - 2019 English: 4th
  - 2020 English: 2nd
  - 2020 Greek: 34th
- It ended up being pretty easy to do hybrid training / ensembling

Related Readings

- OffensEval 2019: Zampieri et al. 2019
- Offenseval 2020: Zampieri et al. 2020
- Greek Data (OGTD): Petnis et al. 2020
- BERT: Devlin et al. 2019
- RoBERTa: Liu et al. 2019
- XLM-RoBERTa: Conneau et al. 2020
- Greek Preprocessing References: Athanasiou et al. 2017

Questions?