Comparative Analysis of Irony and Sarcasm Detection, and Self Attention Approach Development

- **Candidate: Lorenzo Famiglini, 838675**
- **Supervisor: Prof. Elisabetta Fersini**
- **Co-supervisor: Prof. Paolo Rosso**

- **Academic Year 2019-2020**

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Scuola Di Scienze
Dipartimento Di Informatica, Sistemistica E Comunicazione
Corso Di Laurea Magistrale In Data Science

Introduction

"You had one job.

So I waited, 10 minutes and paid 10$ for pre-packaged and bread. Thanks @Starbucks your service today was fantastic!"

Anonymous.

- Research Questions, Objectives and Contributions
- Proposed Methods
- Analyzed studies, Tasks and Dataset
- Results and Conclusions
- Further Studies

Research Questions

- What are the most representative features?
- What is the generalization power of the developed models?
- How encoder output layers can be exploited?
- Can ensemble methods help quality prediction?

Objectives

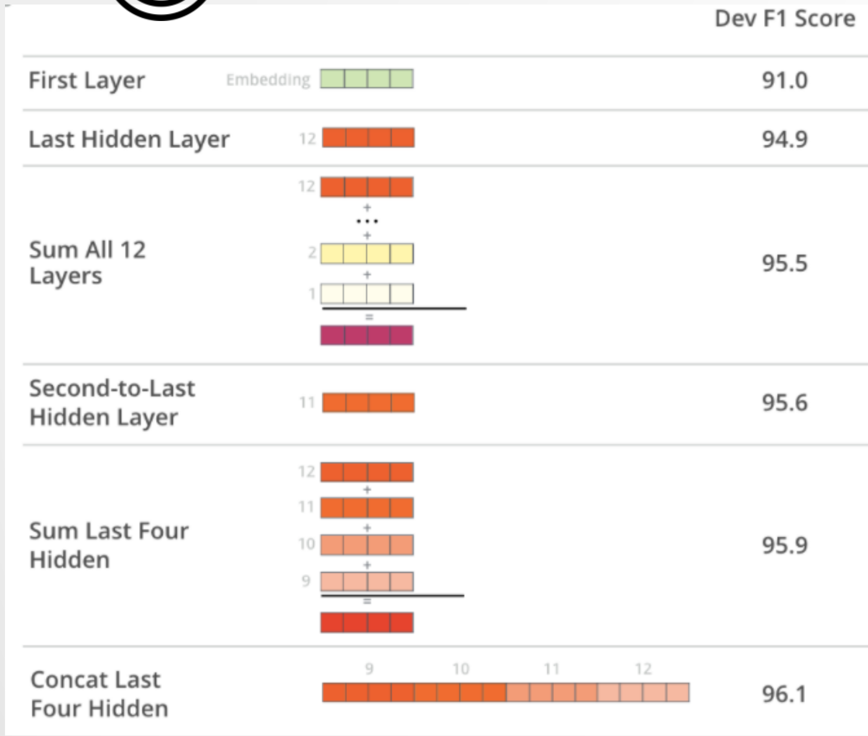
1. Comparison between three different studies
2. Developing new methodologies
3. Understanding models limits within in-domain and out-domain distribution

Contributions

- Comparative study to understand the generalizing power
- Development of a self attention block for creating more contextualized embeddings
- Development of a model based on emotional embeddings by exploiting self attention layer
- Development of the Ensemble of ensembles method

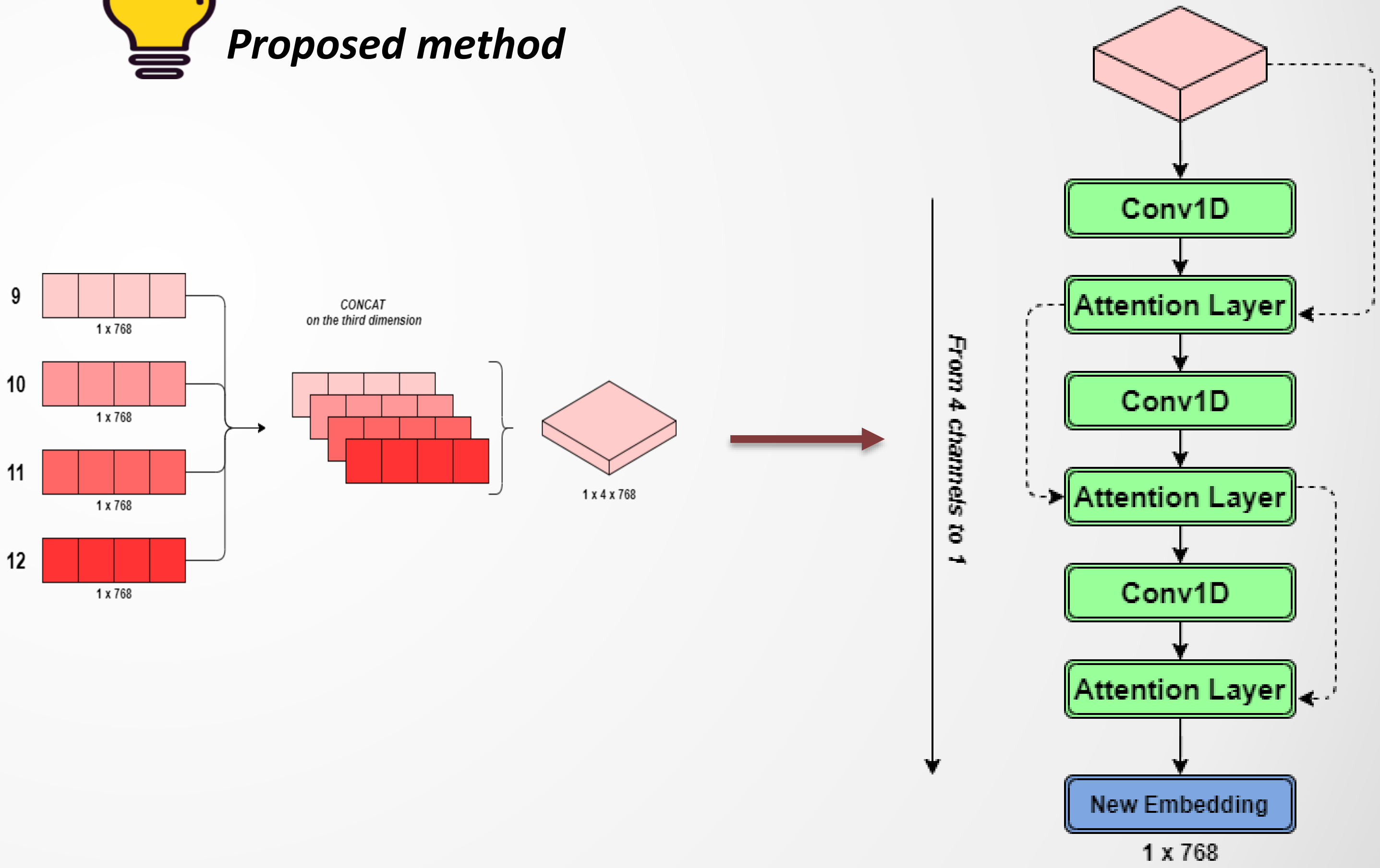
BERTweet Features-based contextualization block

Baseline Approach

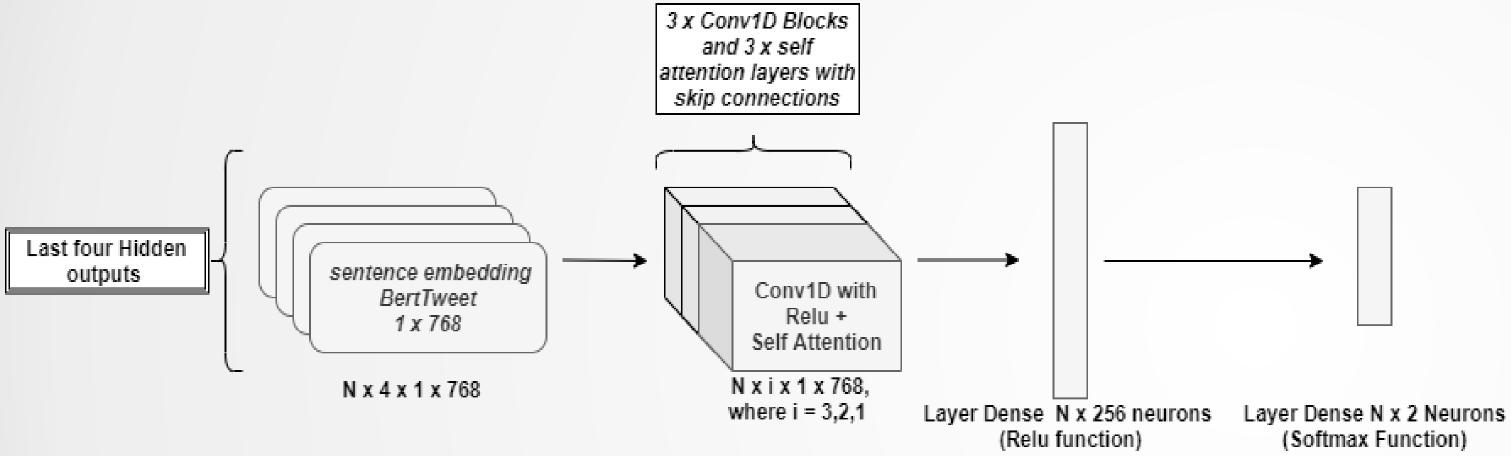


	Dev F1 Score
First Layer	91.0
Last Hidden Layer	94.9
Sum All 12 Layers	95.5
Second-to-Last Hidden Layer	95.6
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1

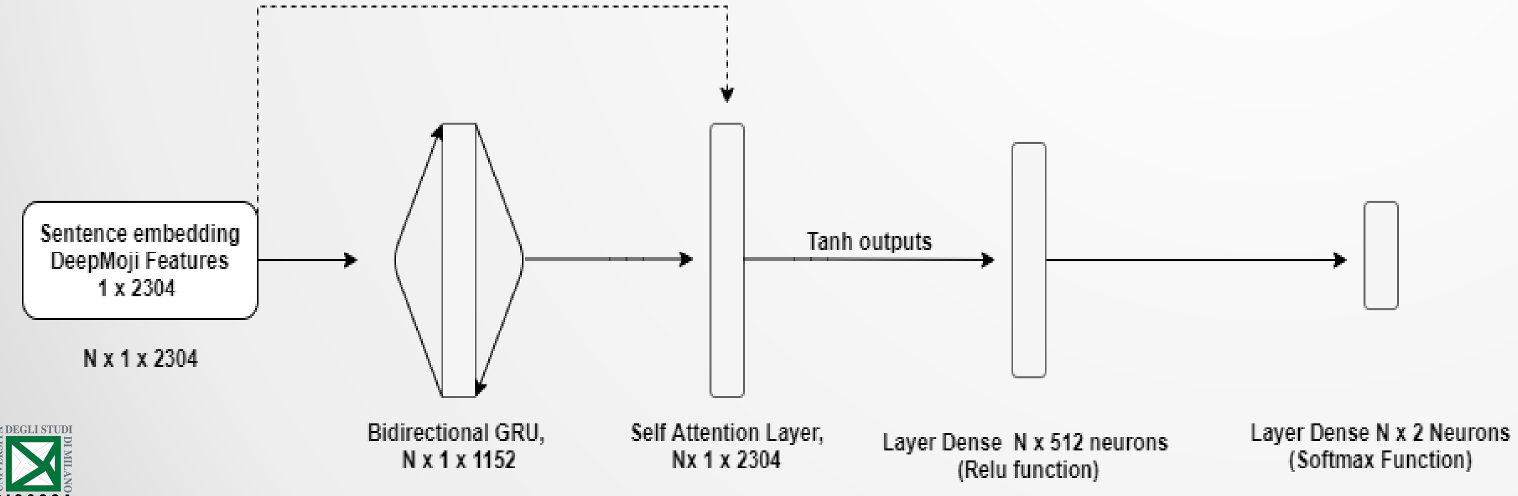
Proposed method



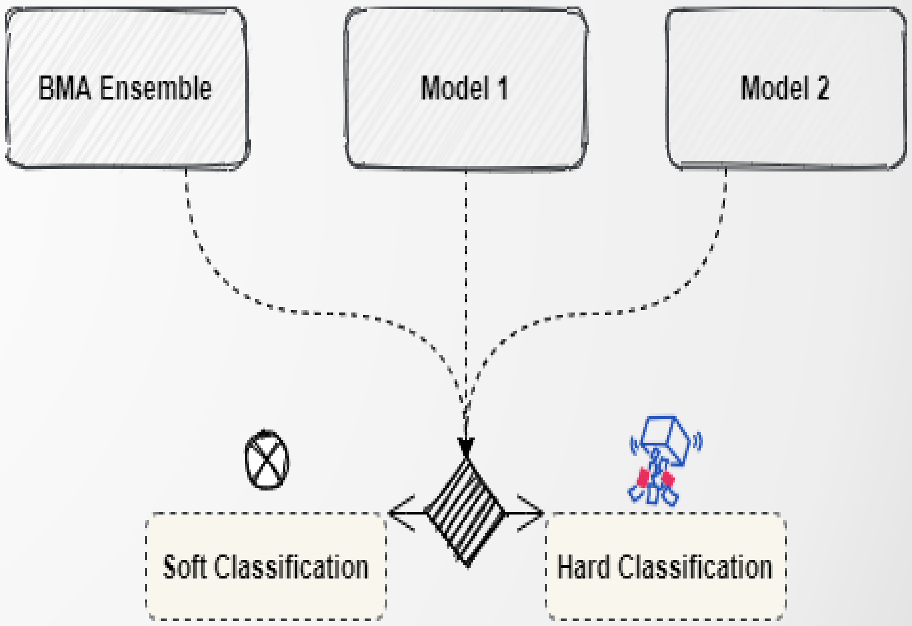
BERTweet Features-based



DeepMoji Features-based



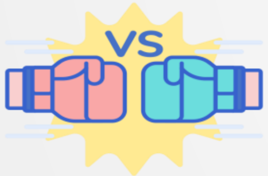
Ensemble of Ensembles



"Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers", Fersini et al. 2015 [1]

Machine Learning models evaluated:

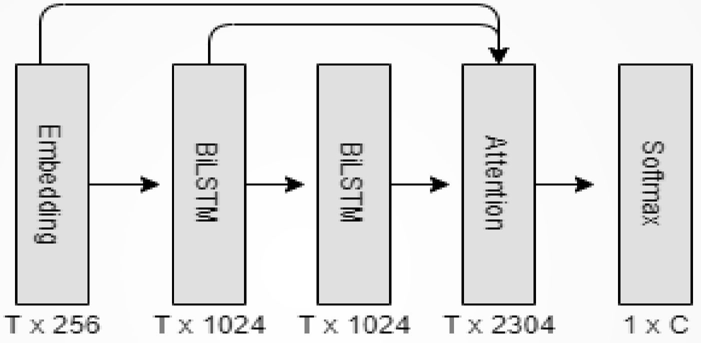
- Logistic Regression
- Extreme Gradient Boosting
- Adaptive Boosting
- Random Forest
- Hist Gradient Boosting



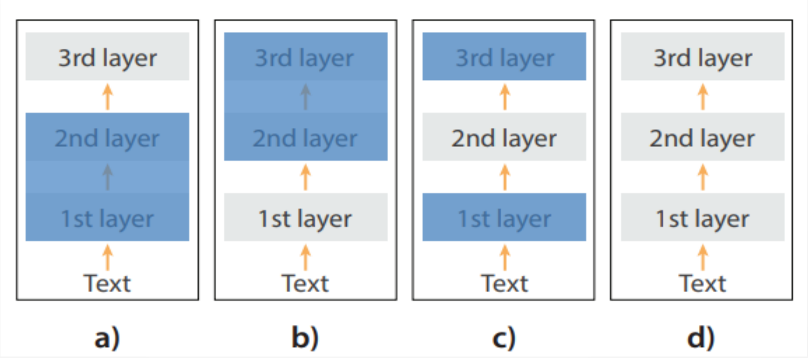
Random Search Bayesian Search[4]
Based on different features combinations

Bayesian Model Averaging ensemble (BMA)

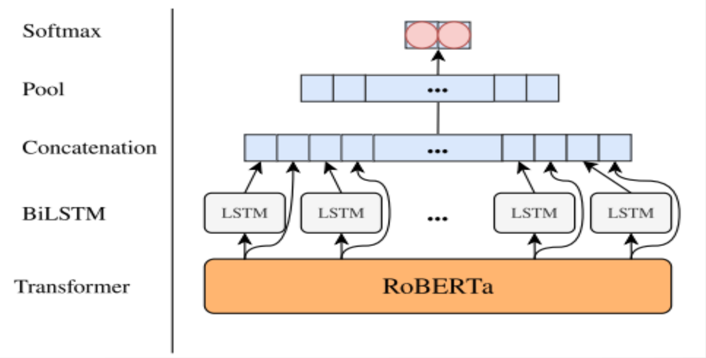
"Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm", Felbo et al. 2017 [2]



Finetuning: Chain-thaw



"A Transformer-based approach to Irony and Sarcasm detection", Potamias et al. 2019 [3]



2 Steps:

- *Features extraction from Roberta*
- *Training the remaining layers*

Sarcasm Detection

Training Set

Source	Sarcasm	Not-Sarcasm
Ptacek et al. 2014 [5]	14.070	16.718
Fersini et al. 2015 [6]	4.000	4.000
Ghosh et al. 2016 [7]	18.488	21.292

Test set

In-Domain distribution:

Ghosh test set [7], 1975 of which 1000 labelled as sarcastic

Out-Domain distribution:

Riloff et al. 2013 [10], 1956 of which 308 labelled as sarcastic

Irony Detection

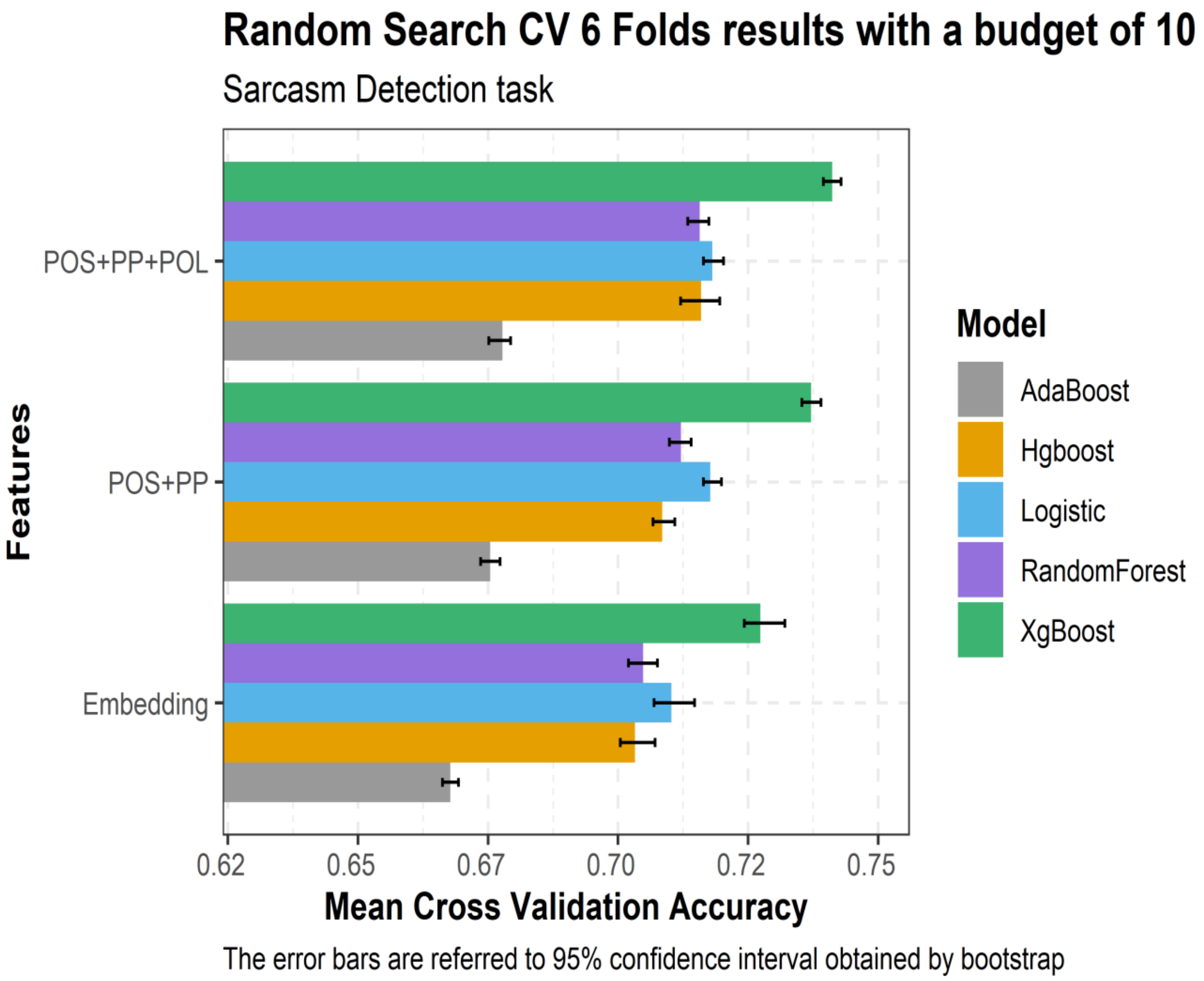
Source	Irony	Not-Irony
SemEval 2018, Task 3A [8]	1898	1904
Reyes-Rosso, 2013 [9]	10.000	30.000

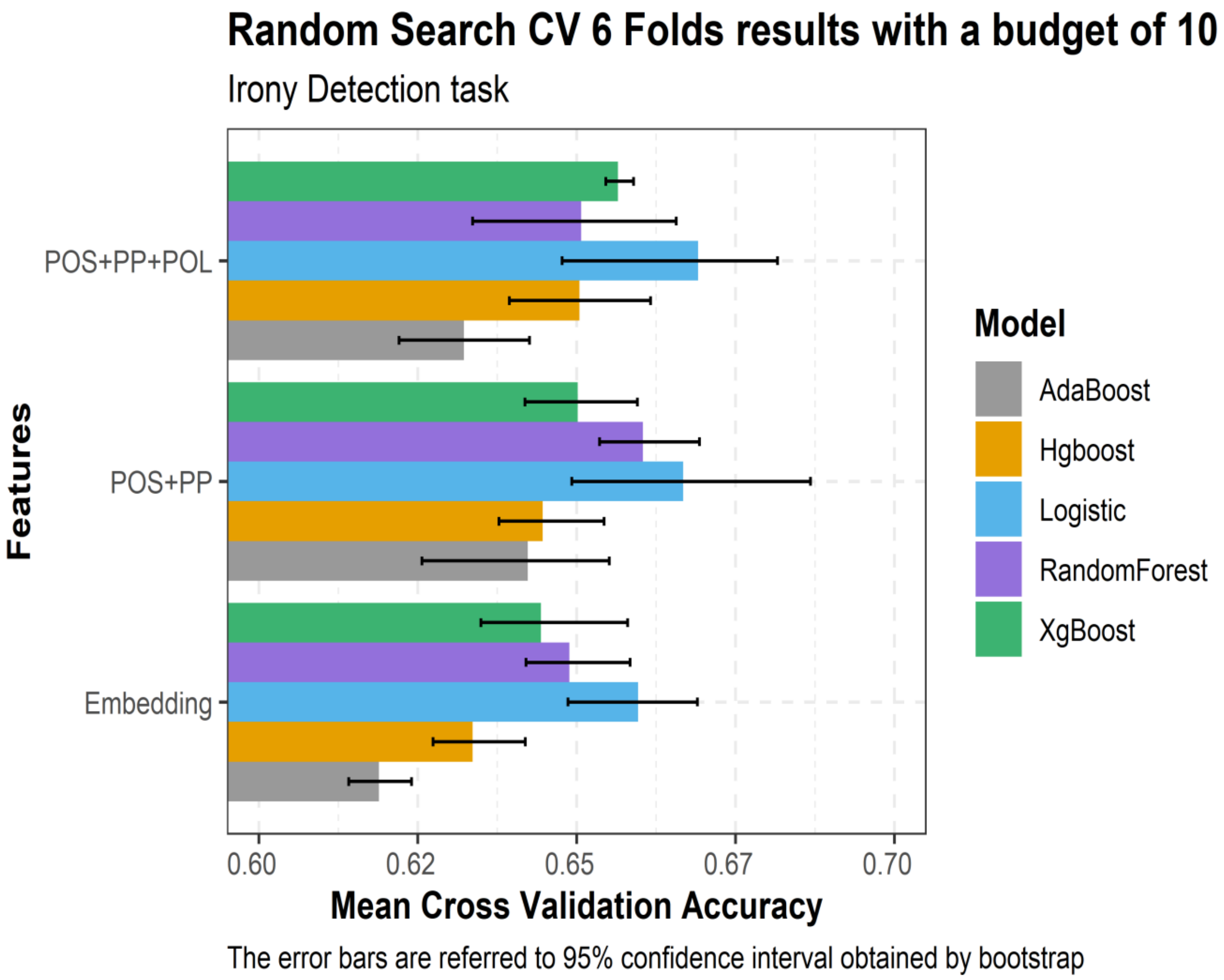
Constrained task

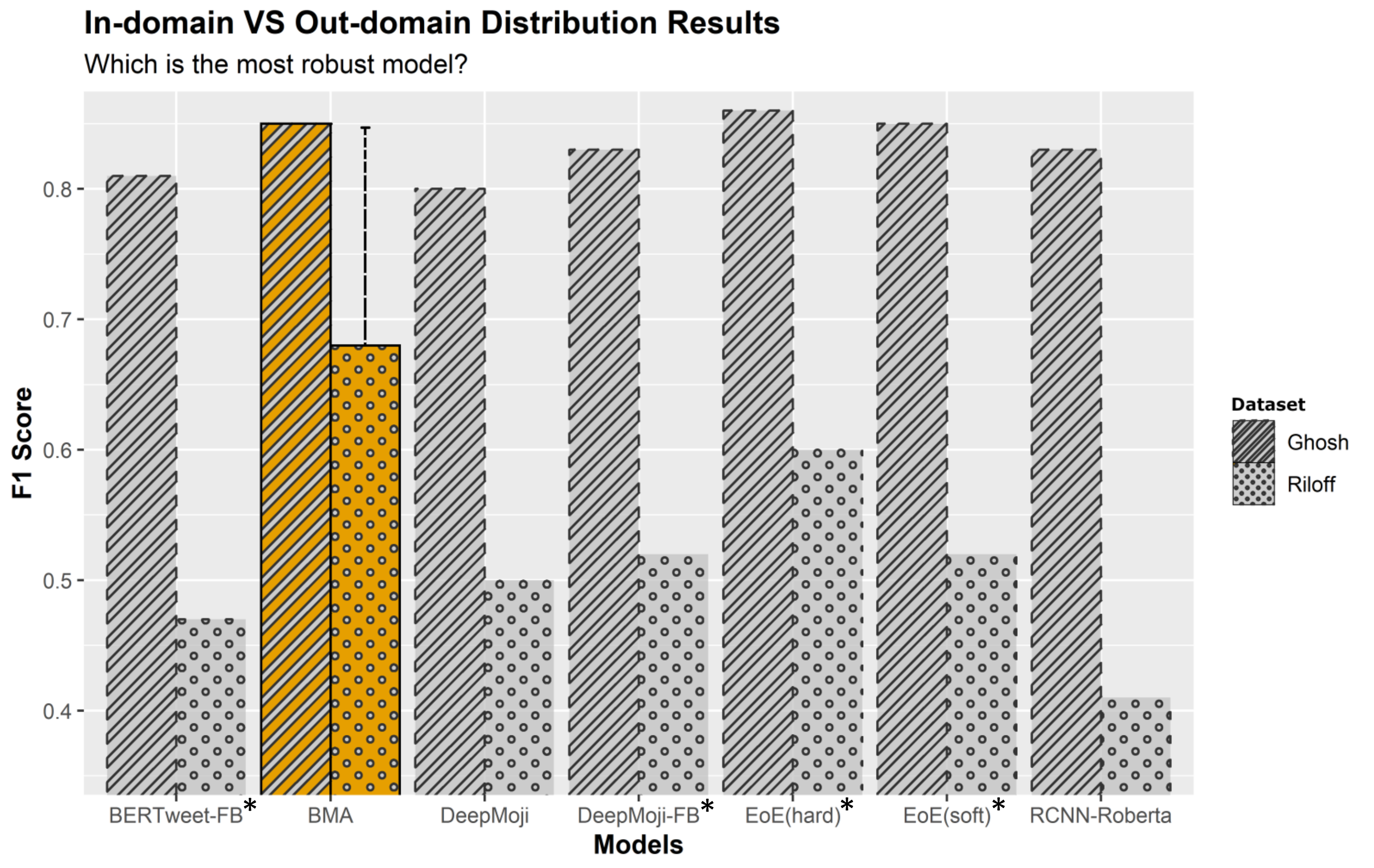
Unconstrained task

In-Domain distribution:

SemEval test set task 3A [8], 784 of which 311 labelled as ironic


Random Search CV 6 Folds results with a budget of 10
Sarcasm Detection task
Features
POS+PP+POL
POS+PP
Embedding
0.62
0.65
0.67
0.70
0.72
0.75
Mean Cross Validation Accuracy
Model
AdaBoost
Hgboost
Logistic
RandomForest
XgBoost
The error bars are referred to 95% confidence interval obtained by bootstrap


Random Search CV 6 Folds results with a budget of 10
Irony Detection task
Features
POS+PP+POL
POS+PP
Embedding
0.60
0.62
0.65
0.67
0.70
Mean Cross Validation Accuracy
Model
AdaBoost
Hgboost
Logistic
RandomForest
XgBoost
The error bars are referred to 95% confidence interval obtained by bootstrap



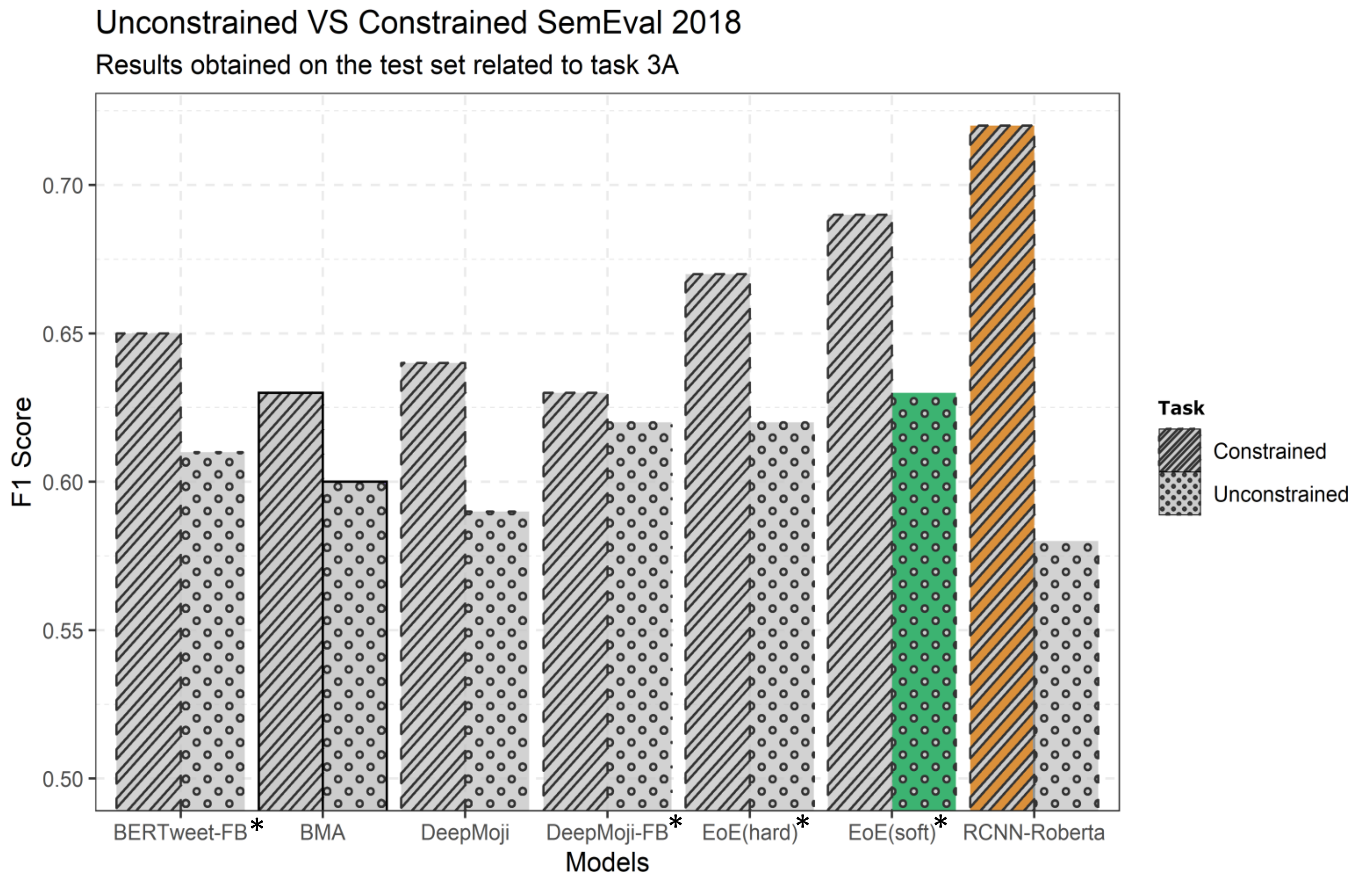
*** New proposed methods within this thesis***

Riloff set:

	Accuracy	F1	Precision	Sensitivity
BERTweet-FB*	0.65	0.47	0.78	0.65
BMA	**0.87**	**0.68**	0.76	**0.90**
DeepMoji	0.71	0.50	**0.80**	0.66
DeepMoji-FB*	0.74	0.52	**0.80**	0.67
EoE(hard)*	0.79	0.60	0.71	0.86
EoE(soft)*	0.72	0.52	0.68	0.82
RCNN-Roberta	0.67	0.41	0.70	0.61

Ghosh test set:

	Accuracy	F1	Precision	Sensitivity
BERTweet-FB*	0.79	0.81	0.79	0.80
BMA	0.84	0.85	0.85	0.84
DeepMoji	0.78	0.80	0.77	0.79
DeepMoji-FB*	0.82	0.83	0.82	0.82
EoE(hard)*	**0.85**	**0.86**	**0.86**	**0.85**
EoE(soft)*	0.83	0.85	0.84	0.83
RCNN-Roberta	0.83	0.83	0.83	0.82



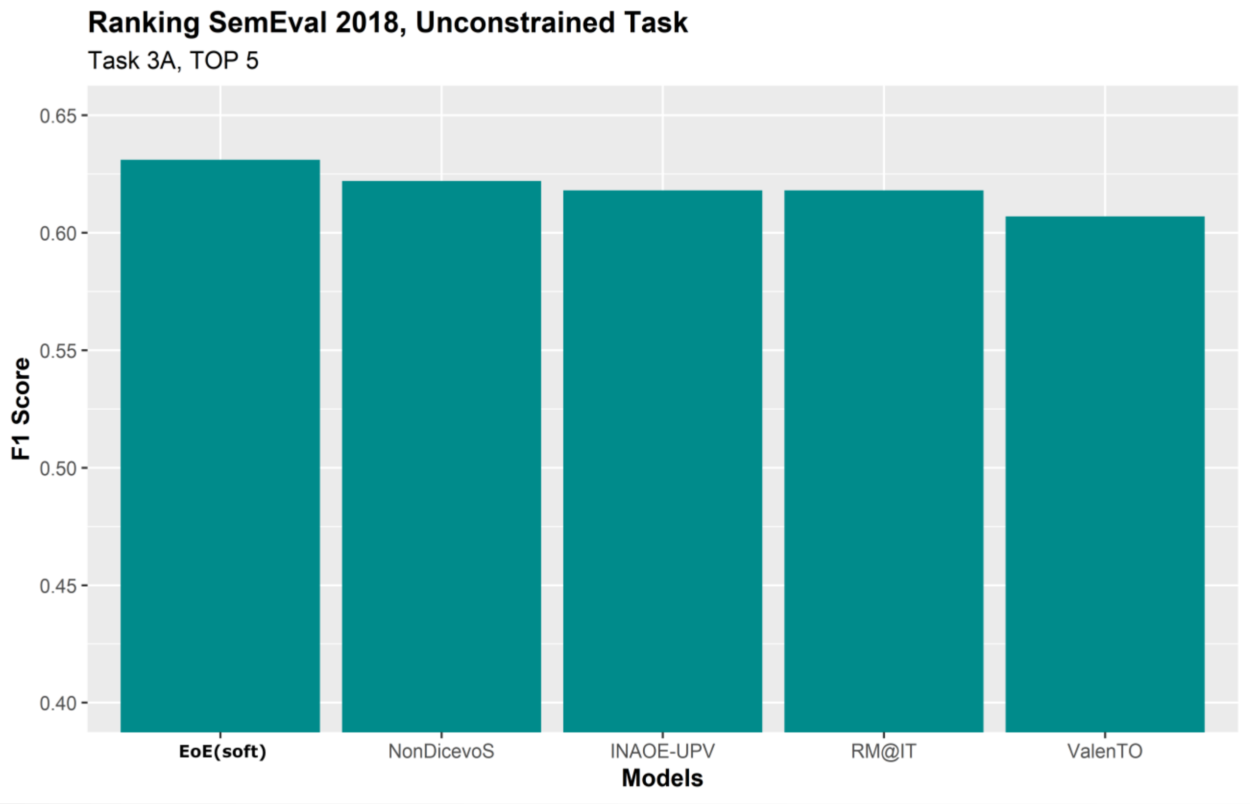
*** New proposed methods within this thesis***

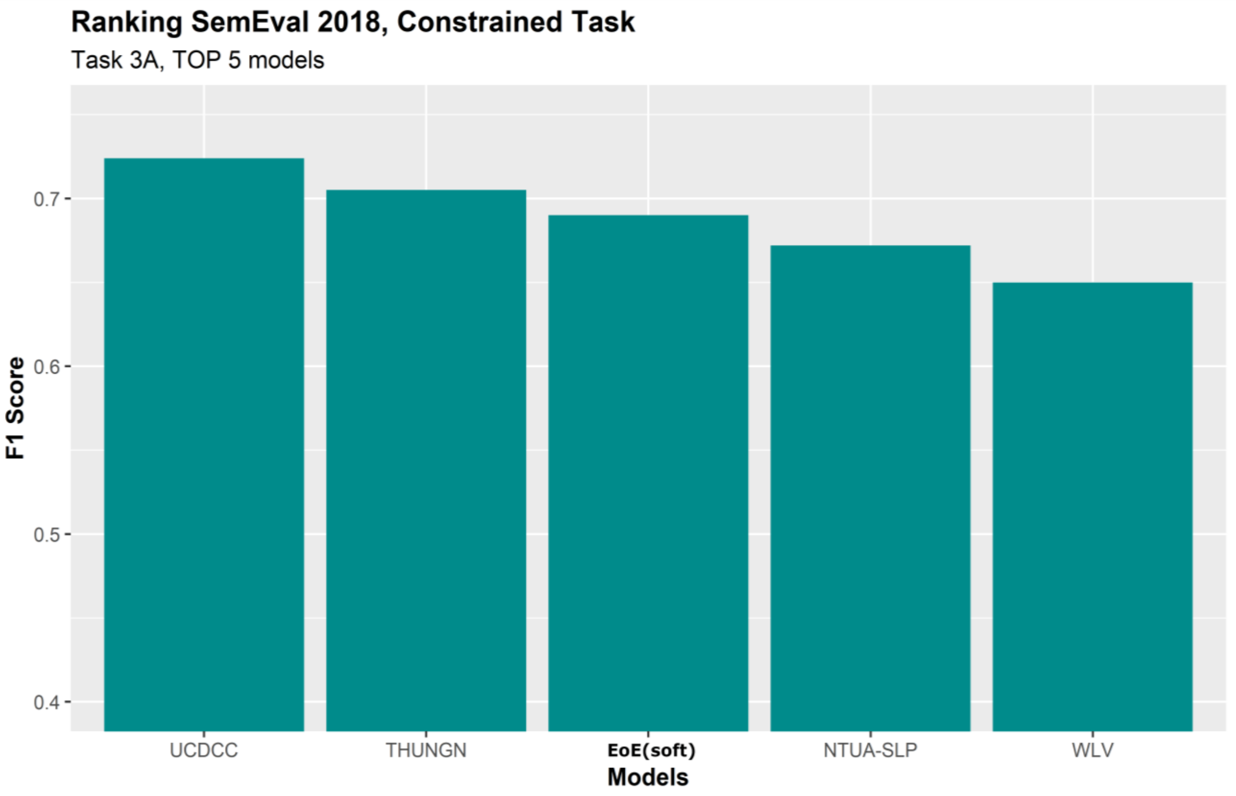
Unconstrained Task:

	Accuracy	F1	Precision	Sensitivity
BERTweet-FB*	0.53	0.61	0.60	0.66
BMA	**0.62**	0.60	0.63	0.64
DeepMoji	0.60	0.59	0.62	0.62
DeepMoji-FB*	0.53	0.61	0.60	**0.66**
EoE(hard)*	0.58	0.62	0.65	0.63
EoE(soft)*	0.61	**0.63**	**0.66**	0.65
RCNN-Roberta	0.59	0.58	0.61	0.61

Constrained Task:

	Accuracy	F1	Precision	Sensitivity
BERTweet-FB*	0.71	0.65	0.71	0.70
BMA	0.69	0.63	0.68	0.66
DeepMoji	0.71	0.64	0.70	0.70
DeepMoji-FB*	0.70	0.63	0.69	0.69
EoE(hard)*	0.73	0.67	0.72	0.72
EoE(soft)*	**0.74**	0.69	0.73	0.74
RCNN-Roberta	**0.74**	**0.72**	**0.76**	**0.75**


Ranking SemEval 2018, Unconstrained Task
Task 3A, TOP 5
F1 Score
0.65
0.60
0.55
0.50
0.45
0.40
EoE(soft)
NonDicevoS
INAOE-UPV
RM@IT
ValenTO
Models


Ranking SemEval 2018, Constrained Task
Task 3A, TOP 5 models
F1 Score
0.7
0.6
0.5
0.4
UCDCC
THUNGN
EoE(soft)
NTUA-SLP
WLV
Models

- Try & Compare different output encoders combination (from BERTweet)
- Analyze word embeddings instead of sentence embeddings
- Add features related to users
- Analyze Reddit Data

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

References:

[1] Enza Messina Elisabetta Fersini Federico Alberto Pozzi. “Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers”. In: (2015)

[2] Anders Søgaard Iyad Rahwan Sune Lehmann Bjarke Felbo Alan Mislove. “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”. In: (2017)

[3] Andreas Georgios Stafylopatis Rolandos Alexandros Potamias Georgios Siolas. “A Transformer-based approach to Irony and Sarcasm detection”. In: (2019)

[4] C K Williams and Rasmussen. Gaussian processes for machine learning. MIT Press, 2006

[5] Ivan Habernal Tomas Ptacek and Jun Hong. “Sarcasm Detection on Czech and English Twitter”. In: (2014). doi: https://www.aclweb.org/anthology/C14- 1022.pdf

[6] Enza Messina Elisabetta Fersini Federico Alberto Pozzi. “Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers”. In: (2015)

[7] T. Veale A. Ghosh. “Fracking Sarcasm using Neural Network”. In: Proceedings of NAACL-HLT (2016)

[8] Els Lefever Cynthia Van Hee and Veronique Host. “SemEval-2018 Task 3: Irony Detection in English Tweets”. In: (2018).

[9] P. Rosso A. Reyes and T. Veale. A multidimensional approach for detecting irony in twitter. 2013.

[10] Prafulla Surve Lalindra De Silva Nathan Gilbert Ruihong Huang Ellen Riloff Ashequl Qadir. “Sarcasm as Contrast between a Positive Sentiment and Negative Situation”. In: (2013)

[11] Davide Buscaldi Antonio Reyes Paolo Rosso. “From humor recognition to irony detection: The figurative languageof social media”. In: Elsevier (2012). doi: https://www.academia.edu/3032533/From_humor_recognition_to_irony_detection_The_figurative_language_of_social_media

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA