Addressing 2016 SENTIPOLC Subtasks with State-Of-The-Art Models

Project and Project Work for the course 'Natural Language Processing'

Giorgio Adragna, Chiara Bellatreccia, Francesco Cavaleri, Marcello Mancino

Dipartimento di Informatica, Scienza e Ingegneria
Alma Mater Studiorum - University of Bologna

February 21, 2025

Overview I

Overview II

Introduction - Problem and Dataset Description

EVALITA
Evaluation of NLP and Speech Tools for Italian

Three subtasks of the SENTIPOLC EvalITA 2016 challenge [Barbieri et al., 2016]:

- **Irony Detection** - Project
- **Sentiment Analysis** - Project Work
- **Subjectivity Detection** - Project Work

The dataset of the official challenge consists of over Italian 9000 tweets from 2016, mostly political.

Introduction - Problem and Dataset Description

A total of **six binary labels:**

- one **iro** label for irony detection
- four **opos, oneg, lpos, lneg** labels for sentiment analysis
- one **subj** label for subjectivity detection.
  An objective tweet has always opos, oneg, lpos, lneg all equal to 0.

subj	opos	oneg	iro	lpos	lneg	**description and *explanatory tweet in Italian***
0	0	0	0	0	0	objective *l'articolo di Roberto Ciccarelli dal manifesto di oggi* `http://fb.me/1BQVy5WAk`
1	0	0	0	0	0	subjective with neutral polarity and no irony *Primo passaggio alla #strabrollo ma secondo me non era un iscritto*
1	1	0	0	1	0	subjective with positive polarity and no irony *splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura* `http://t.co/GWoZqbxAuS`
1	0	1	0	0	1	subjective with negative polarity and no irony *Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont...* `http://t.co/3CazKS7Y`
1	1	1	0	1	1	subjective with both positive and negative polarity (mixed polarity) and no irony *Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"* `http://t.co/kIKnbFY7`
1	1	0	1	1	0	subjective with positive polarity, and an ironic twist *Questo governo Monti dei paschi di Siena sta cominciando a carburare; speriamo bene...*
1	1	0	1	0	1	subjective with positive polarity, an ironic twist, and negative literal polarity *Non riesco a trovare nani e ballerine nel governo Monti. Ci deve essere un errore! :)*
1	0	1	1	0	1	subjective with negative polarity, and an ironic twist *Calderoli: Governo Monti? Banda Bassotti ..infatti loro erano quelli della Magliana.. #FullMonti #fuoritutti #piazzapulita*
1	0	1	1	1	0	subjective with negative polarity, an ironic twist, and positive literal polarity *Ho molta fiducia nel nuovo Governo Monti. Più o meno la stessa che ripongo in mia madre che tenta di inviare un'email.*
1	1	0	1	0	0	subjective with positive polarity, an ironic twist, and neutral literal polarity *Il vecchio governo paragonato al governo #monti sembra il cast di un film di lino banfi e Renzo montagnani rispetto ad uno di scorsese*
1	0	1	1	0	0	subjective with negative polarity, an ironic twist, and neutral literal polarity *arriva Mario #Monti: pronti a mettere tutti il grembiulino?*
1	1	0	1	1	1	subjective with positive polarity, an ironic twist, and mixed literal polarity *Non aspettare che il Governo Monti prenda anche i tuoi regali di Natale... Corri da noi, e potrai trovare IDEE REGALO a partire da 10e...*
1	0	1	1	1	1	subjective with negative polarity, an ironic twist, and mixed literal polarity *applauso freddissimo al Senato per Mario Monti. Ottimo.*

Introduction - Previous Approaches

Previous
approaches to the 2016 SENTIPOLC challenge included:

- **Machine Learning-based tools** [Di Rosa and Durante, 2016]
- **Feature-based models** [Buscaldi and Farías, 2016]
- **SVM + feature extraction** [Passaro et al., 2016]

In general, in 2016 PLMs and Transformer-based models were yet to be explored. For example, the BERT paper was written in 2019 [Devlin et al., 2019].

System	Non-Iro	Iro	F
tweet2check16.c	0.9115	0.1710	0.5412
CoMoDI.c	0.8993	0.1509	0.5251
tweet2check14.c	0.9166	0.1159	0.5162
IRADABE.2.c	0.9241	0.1026	0.5133
ItaliaNLP.1.c	0.9359	0.0625	0.4992
ADAPT.c	0.8042	**0.1879**	0.4961
IRADABE.1.c	0.9259	0.0484	0.4872
Unitor.2.u	**0.9372**	**0.0248**	**0.4810**
Unitor.c	0.9358	0.0163	0.4761
Unitor.1.u	0.9373	0.0084	0.4728
ItaliaNLP.2.c	**0.9367**	0.0083	0.4725
Baseline	0.9376	0.000	0.4688

Our Approach

GOALS

Irony Detection: enhance (positive) irony detection experimenting with different configurations

Sentiment Analysis and Subjectivity Detection: check if the correlation between these two tasks can enhance classification

Our Approach for Irony Detection - Architectures

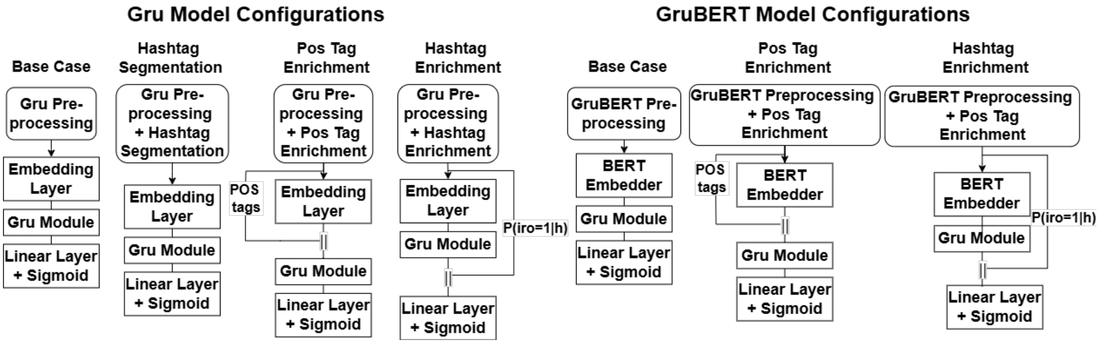
Gru Model

A Gru-based architecture with the possibility of incorporating *Pos Tag Enrichment* and *Hashtag Enrichment*.

GruBERT Model [Horne et al., 2020]

A BERT-based architecture using AlBERTo [Polignano et al., 2019] with a Gru module, and the possibility of incorporating *Pos Tags Enrichment* and *Hashtag Enrichment*.

Our Approach for Irony Detection - Architectures



Our Approach for Irony Detection - Data Preprocessing

Gru Model Preprocessing

- substitute emoticons and emojis with (translated) textual descriptions
- lowercasing
- replace URLs with token
- remove mentions, HTML expressions, retweet markers, unnecessary duplicate letters, redundant symbols
- expanding abbreviations

GruBERT Preprocessing

- no substitution of emoticons and emojis, because it uses AlBERTo
- no abbreviations expansion and no removal of duplicate characters

Additional Preprocessing

+ POS Tags Enrichment, Hashtag Enrichment and Hashtag Segmentation when needed

Our Approach for Irony Detection - Experimental Setup

Grid Search using the Gru Model for 30 epochs, focusing on:

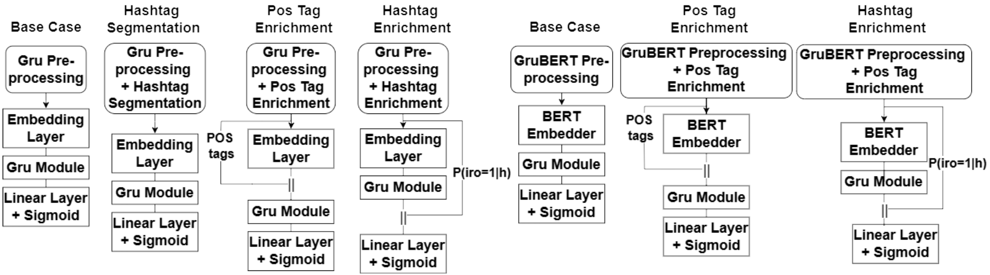
- Loss function
- Learning rate
- Label smoothing
- Number of Gru layers
- Gru dropout probability

Gru Model final training

Trained in four different configurations:
Base Case, Hahtag Segmentation, Pos Tags Enrichment, Hashtag Enrichment

GruBERT Model final training

Trained in three different configurations:
Base Case, Pos Tags Enrichment, Hashtag Enrichment



Our Approach for Sentiment Analysis and Subjectivity Detection - Architectures

BERT Baseline
A BERT-based Baseline which outputs all the five labels in parallel

BERT SubjSent
A BERT-based architecture which uses subjectivity detection to influence sentiment analysis

Formula

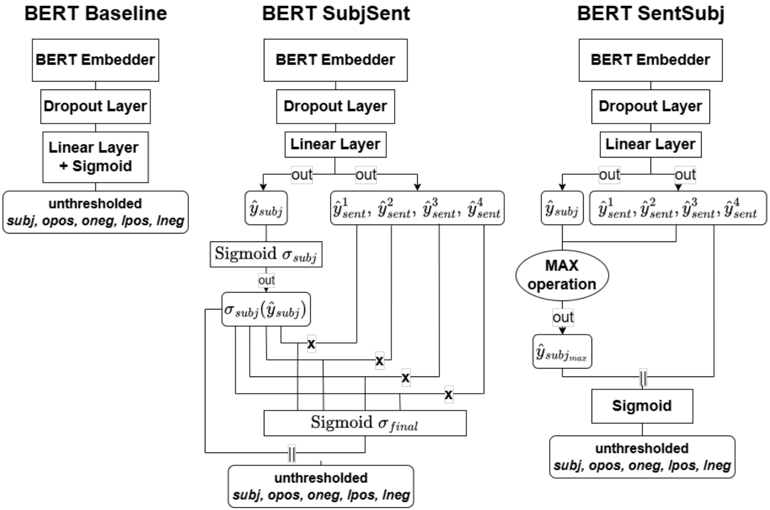
$s_{sent} = \sigma_{final}(\sigma_{subj}(\hat{y}_{subj}) \cdot \hat{y}_{sent}^{(i)})$
$i = 1 \ldots 4$

BERT SentSubj
A BERT-based architecture which uses sentiment analysis to influence subjectivity detection

Formula

$\hat{y}_{subj_{max}} =$
$\max(\hat{y}_{subj}, \hat{y}_{sent}^{(1)}, \hat{y}_{sent}^{(2)}, \hat{y}_{sent}^{(3)}, \hat{y}_{sent}^{(4)})$

Our Approach for Sentiment Analysis and Subjectivity Detection - Architectures



Our Approach for Sentiment Analysis and Subjectivity Detection - Data Preprocessing and Experimental Setup

Preprocessing to remove URLs, mentions, HTML expressions, retweet markers and to replace abbreviations

Grid Search on the BERT backbone using the BERT Baseline Model for 30 epochs (10 + 20), focusing on:

- Batch size
- Label smoothing
- Weight decay
- Last layer bias initialization

Final Training of the three models employing the BERT backbone from the Baseline

Results and Discussion

Results and Discussion - Irony Detection

- Both the models in any configuration surpass the dummy baseline results AND the 2016 challenge results, **except** for the $F_{iro=0}$ score, for which the dummy majority baseline remains unmatched $\longrightarrow$ dataset imbalance
- The **GruBERT** model in its **Hashtag Enrichment** configuration results to be the best model to maximize irony detection, with $F_{iro=1} = 0.437$
  $\longrightarrow$ maximum $F_{iro=1}$ in 2016: 0.1879... but without PLMs! Seems fair.
- The model has trouble in distinguishing between *literal* and *intended* meaning (i.e. the inversion of polarity that is irony) and has a **lack of contextual understanding**

	Test Set	
	$F_{iro=0}$	$F_{iro=1}$
Dummy Majority	0.934	0.000
2016 SENTIPOLC Challenge	0.937	0.188
GruBERT + Hashtag Enrichment	0.910	0.444

Results and Discussion - Sentiment Analysis and Subjectivity Detection

- The best performing model on the Test Set is the **BERT SubjSent model**, with $F_{subj} = 0.743$, $F_{overall} = 0.675$ and $F_{literal} = 0.666$, surpassing the BERT Baseline and BERT SentSubj in **all** the labels
- The **BERT SubjSent** architecture reaches the goal of (slightly) enhancing the sentiment analysis performance using the subjectivity detection, while for the **BERT SentSubj** the vice versa does not happen
- The model struggles to differentiate **literal** polarity from **overall** polarity
- **Strong correlation between subjectivity and polarity errors**

	Test Set		
	F_{subj}	$F_{overall}$	$F_{literal}$
BERT Baseline	0.736	0.651	0.660
BERT SubjSent	0.743	0.675	0.666

Possible Improvements

As for **Irony Detection**, discern **literal** and **intended** meaning [Yi and Xia, 2025] and integrate **contextual data** [Helal et al., 2024, Wallace et al., 2015]

As for **Sentiment Analysis** and **Subjectivity Detection**, employ a **two-stage approach** to enhance the correlation between the labels and use **contrastive learning** [Wang et al., 2023]

Thanks for your attention!

APPENDIX: Examples

Irony Detection

"mario monti? preferisco capperi e acciughe! (il trota)" → **False Positive**
"mario monti è più forte di chuck norris #monti #ministri" → **False Positive**
"281 senatori a favore del governo tecnico. mario monti è pronto ad instaurare il reich" → **False Negative**
"mario monti nominato europeo dell'anno. e gli altri chi erano? hitler e stalin?" → **False Negative**

Sentiment Analysis and Subjectivity Detection

"Ottimo lavoro! Il governo ha di nuovo fallito." → **False opos Positive**
"Questa legge ha lati positivi e negativi." → **False Ineg Positive**
"Non è così terribile come sembra." → **False Ineg Positive**

APPENDIX: Dimensionality Reduction

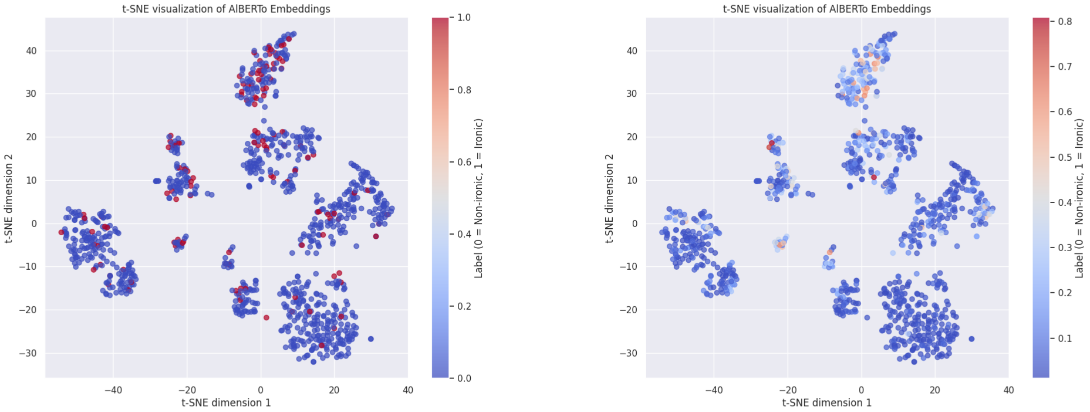


Figure: Ground truth and prediction t-SNE plots of BERT embeddings for irony detection.

APPENDIX: Full Results - Irony Detection

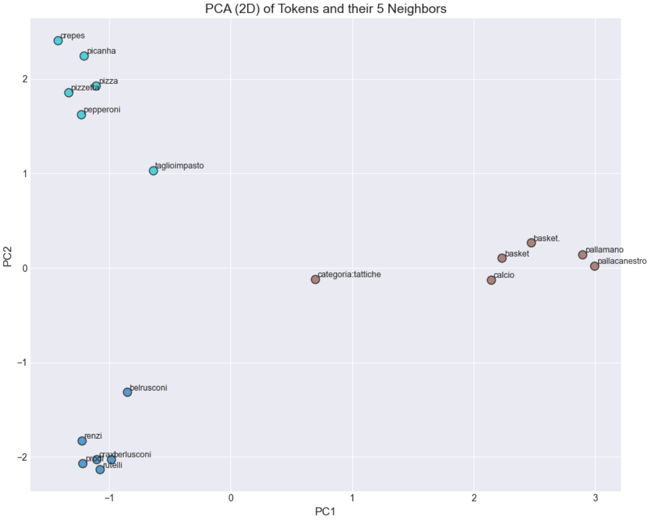
		Validation Set			Test Set		
		$F_{iro=0}$	$F_{iro=1}$	F_{iro}	$F_{iro=0}$	$F_{iro=1}$	F_{iro}
dummy baselines	**random**	0.646	0.172	0.409	0.644	0.198	0.421
	majority	0.943	0.000	0.471	0.934	0.000	0.467
Gru Model	**base case**	0.920	0.400	0.660	0.916	0.418	0.667
	hashtag enrichment	0.921	0.398	0.660	0.918	0.415	0.667
	hashtag segmentation	0.926	0.385	0.655	0.919	0.380	0.650
	pos tags enrichment	0.903	0.351	0.627	0.908	0.403	0.656

		Validation Set			Test Set		
		$F_{iro=0}$	$F_{iro=1}$	F_{iro}	$F_{iro=0}$	$F_{iro=1}$	F_{iro}
dummy baselines	**random**	0.646	0.172	0.409	0.644	0.198	0.421
	majority	0.943	0.000	0.471	0.934	0.000	0.467
GruBERT Model	**base case**	0.921	0.367	0.644	0.916	0.432	0.674
	hashtag enrichment	0.914	0.370	0.642	0.910	0.444	0.677
	pos tags enrichment	0.920	0.362	0.641	0.917	0.428	0.672

APPENDIX: Full Results - Sentiment Analysis and Subjectivity Detection

	Validation Set				Test Set			
	F_{subj}	$F_{overall}$	$F_{literal}$	F_{score}	F_{subj}	$F_{overall}$	$F_{literal}$	F_{score}
random	0.470	0.457	0.482	0.469	0.508	0.462	0.481	0.484
majority	0.387	0.426	0.430	0.414	0.402	0.405	0.412	0.407
BERT Baseline	0.776	0.746	0.736	0.753	0.736	0.651	0.660	0.682
BERT SubjSent	0.773	0.747	0.737	0.752	0.743	0.675	0.666	0.695
BERT SentSubj	0.777	0.746	0.737	0.753	0.734	0.649	0.661	0.681

APPENDIX: PCA of Tokens and their 5 Neighbors



References I

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016).
Overview of the evalita 2016 sentiment polarity classification task.
In Cignarella, A. T., Bosco, C., Mazzei, A., and Tamburini, F., editors, *EVALITA. Evaluation of NLP and Speech Tools for Italian*. Accademia University Press.

Buscaldi, D. and Farías, D. I. H. (2016).
Iradabe2: Lexicon merging and positional features for sentiment analysis in italian.
In *CLiC-it/EVALITA*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
Bert: Pre-training of deep bidirectional transformers for language understanding.

Di Rosa, E. and Durante, A. (2016).
Tweet2Check evaluation at Evalita Sentipolc 2016, pages 189–193.

Helal, N., Abdelgawad, A., Badr, N., and Afify, Y. (2024).
A contextual-based approach for sarcasm detection.
Scientific Reports, 14.

References II

Horne, L., Matti, M., Pourjafar, P., and Wang, Z. (2020).
GRUBERT: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis.
In Shmueli, B. and Huang, Y. J., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 130–138, Suzhou, China. Association for Computational Linguistics.

Passaro, L. C., Bondielli, A., and Lenci, A. (2016).
Exploiting emotive features for the sentiment polarity classification of tweets.
In *CLiC-it/EVALITA*.

Polignano, M., Basile, P., Degemmis, M., Semeraro, G., and Basile, V. (2019).
Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets.
In *Italian Conference on Computational Linguistics*.

References III

Wallace, B. C., Choe, D. K., and Charniak, E. (2015).
Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment.
In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Wang, X., Zhang, M., Chen, B., Wei, D., and Shao, Y. (2023).
Dynamic weighted multitask learning and contrastive learning for multimodal sentiment analysis.
Electronics, 12(13).

Yi, P. and Xia, Y. (2025).
Irony detection, reasoning and understanding in zero-shot learning.