On the evaluation of BERT-based models for Italian Sentiment Analysis

Steve Azzolin
University of Trento
221251
steve.azzolin@studenti.unitn.it

Abstract

This document is the report for the final Natural Language Understanding course project. The goal was to test several pretrained BERT-based models on different Italian datasets from different domains, with and without fine-tuning, in order to assess their generalization capabilities. An additional online supporting material (SOM) is provided in order to aid the evaluation, with detailed per-class macro and weighted F1 scores[1]

1 Introduction

Sentiment Analysis is a common text categorization task that involves the extraction of sentiment, i.e, positive, neutral, or negative orientation that a writer expresses toward some aspects.

Historically, Sentiment Analysis had been addressed by approaches like Term Counting (Kennedy and Inkpen, 2006) or lexicon-based (Hutto and Gilbert, 2015), in which is available a list of words labelled according to their semantic orientation as either positive or negative (the sentiment lexicon). One big disadvantage of them is that building such lexicons is complex and time-consuming, often involving qualified annotators. In recent times, new developments in Deep Learning architectures have allowed the pre-training of very large models in an unsupervised manner, directly on web data. It has been shown that such models are capable of addressing a large number of different NLP tasks in few-shot learning settings (Vaswani et al., 2017) (Devlin et al., 2019) achieving new state-of-the-art results in many of them.

This work presents an extensive comparison between two BERT based models, namely AlBERTo (Polignano et al., 2019) and Feel-IT (Bianchi et al., 2021), trained on different Italian datasets coming from different domains (Twitter, Product reviews, Hotel reviews, psychotherapy dialogues, Italian Classic Opera) in order to assess their intra-domain and inter-domain generalization capabilities.

[1] https://bit.ly/3Bqorg7

2 Datasets

In this section a brief summary of all datasets is provided, along with some basic statistics in Table (1) and some examples in Table (2). For a detailed description of the file format, please visit the respective web pages. Datasets without an official train-test split were manually splitted in a stratified manner.

SENTIPOLC16[2] is a collection of socio-politcal Italian Tweets, labelled with subjectivity, sentiment polarity, literal sentiment polarity and irony. Sentiment polarity, in addition to neutral, positive, and negative, includes also a mixed class, which will be discarded in our experiments since only another dataset have such label.

Feel-IT[3] contains Italian Tweets collected between 20th August to 12th October 2020 by monitoring trending topics each day. Neutral tweets were removed by the authors in order to have just positive and negative labels as an outcome.

MultiEmotions-IT[4] contains manually annotated comments to music videos and advertisements posted on YouTube and Facebook. As for SENTIPOLC16, positive and negative polarities are not mutually exclusive: a comment can have a mixed polarity containing both positive and negative opinions on different aspects of the media content.

Amazon reviews is a collection of product's reviews automatically collected via the *Amazon Re-*

[2] http://www.di.unito.it/~tutreeb/sentipolc-evalita16/index.html

[3] https://github.com/MilaNLProc/feel-it

[4] https://github.com/RacheleSprugnoli/Esercitazioni_SA/tree/master/dataset

view Export Google Chrome extension[5]. Reviews with less and more than 3 stars are labelled respectively with negative and positive polarity, whereas reviews with 3 stars are labelled as neutral. Note, however, that the lack of direct supervision in the labelling process led to a potentially noisy dataset, in particular for the neutral class, as can be seen from some bad neutral examples illustrated in Table (4).

Coadapt valence is a collection of Italian ABC notes collected by a Personal Healthcare Agent that were annotated with valence by human annotators at Functional Unit level, in the context of a Digital Cognitive Behavioral Therapy (DCBT) intervention (Mousavi et al., 2021). For the task of Sentiment Analysis, each sample is labelled with a label $l \in \{negative, positive, neutral\}$. The dataset comes with a training split and four partitions: partitions A and B are chosen as validation split, while C and D as test split.

Trip-maml[6] was originally intended as a Multi-Aspect Multi-Lingual dataset for aspect-oriented opinion mining, consisting of Tripadvisor hotel reviews in English, Italian, and Spanish. Since our work deals with span-level sentiment analysis, the overall rating of the review (that is an integer value in the range [1,5]) is taken as ground truth, in a similar manner as the review's stars in Amazon reviews.

AriEmozione 1.0[7] (Ari) contains a selection of 678 operas composed between 1655 and 1765 written in the 18th century Italian. Each single verse is annotated with an emotion in the set {love, joy, admiration, anger, sadness, fear, none}, along with the confidence of annotation (strong doubts, quite sure, totally sure). In order to conduct Sentiment Polarity classification, these emotions are compacted in the following way: {love, joy, admiration} → positive, {anger, sadness, fear} → negative, and {none} → neutral.

Dataset	#	pos	neg	neutral
SentiPolc16 train	6970	23	36	41
SentiPolc16 test	1964	16	37	47
Feel-IT	2037	36	64	-
Amazon train	820	68	22	10
Amazon test	352	68	21	11
Multi.E. train	2533	64	25	11
Multi.E. test	447	64	25	11
Coadapt train	3417	13	27	60
Coadapt test	439	12	28	60
Coadapt val	417	14	26	60
Trip-maml train	292	72	12	16
Trip-maml test	125	70	14	16
Ari train	1962	43	55	2
Ari test	250	43	53	4

Table 1: Summary statistics for all datasets: name, number of samples, percentage of positive, negative and neutral samples respectively

3 Preprocessing

Preprocessing is one of the most important steps in a NLP pipeline. I followed the approach of (Polignano et al., 2019) of using Ekphrasis (Baziotis et al., 2017) for text normalization and word segmentation (for splitting hashtags). Take the following example:

> @matteorenzi @MiurSocial #labuonascuola basta supplenti ma anche coraggio nel valutare i prof https://t.co/mnlTnLMzpT

The resulting processed text is:

> <user> <user> <hashtag> la buona scuola </hashtag> basta supplenti ma anche coraggio nel valutare i prof <url>

The SentencePiece tokenizer[8] is then used to split tweets into a list of tokens, which will be later converted into vocabulary indexes. SentencePiece is a language-independent sub-word tokenizer and detokenizer designed for Neural-based text processing that creates sub-word units specifically to the size of the chosen vocabulary and the language of the corpus. In this way, words that are not in the vocabulary are decomposed into sub-words and character tokens that we can then generate embeddings for.

4 Models

4.1 AlBERTo

AlBERTo (Polignano et al., 2019) is a BERT based model which was the first Italian language understanding model to represent the social media language, Twitter in particular. It has been trained on

[5] https://chrome.google.com/webstore/detail/amazon-review-export/ikphihiljfhlmpokjbmkhliphckfpcph

[6] https://github.com/diegma/trip-maml

[7] https://github.com/TinfFoil/AriEmozione

[8] https://github.com/google/sentencepiece

Dataset	Text	label
SentiPolc16	Bossi risponde con una pernacchia a un ipotetico governo Monti e con il dito medio a misure destinate alle pensioni. Un Signor ministro...	negative
Feel-IT	Elisa ribelle del mio cuore 💗 #elisadirivombrosa	positive
Amazon	Si è fuso subito: bello, ha tutto, mandrino autoserrante molto comodo: 4 stelleal primo lavoro un po' piu' gravoso si è fuso il motore : 1 stellaassistenza sovrumana: fuso lunedì tardo pomeriggio ricevuto il nuovo martedì nel pomeriggio: 5 stellevalutazione complessiva: 3 stelle	neutral
Multi.E.	Sembra "Ragazzo Inadeguato" di Max Pezzali	neutral
Coadapt	vedo mio figlio arrabbiato e non vuole parlarne	negative
Trip-maml	Posto isolato molto démodé moquette lisa ed arredi anni 70 mal tenuti . Personale quasi tutto di colore , improbabile e di scarsissima professionalità . . manca il ristorante e la colazione è di basso livello . . evitate gente ! !	neutral
Ari	Infelice e sventurato potrà farmi ingiusto fato ma infedele io non sarò	positive

Table 2: Examples for each dataset

TWITA (Basile et al., 2018), a 191GB sized dataset of raw tweets collected from 2012 to 2020. After the typical BERT unsupervised training, AlBERTo is extended with a non-linear classification head, on top of the [CLS] embedding. The whole network is then fine-tuned end-to-end on SENTIPOLC16. AlBERTo uses the SentencePiece tokenizer with a vocabulary size of 128000 and with a max sequence length of 128 tokens, which will be maintained trough all the experiments, also for the following model.

4.2 Feel-IT

The model presented in (Bianchi et al., 2021) uses the Italian BERT model UmBERTo trained on Commoncrawl-ITA[9] for a total of 69GB of raw data, and fine-tuned on the Feel-IT dataset. UmBERTo inherits from the RoBERTa (Liu et al., 2019) model architecture which improves the initial BERT by identifying key hyper-parameters for better results. Umberto extends RoBERTa in two ways: SentencePiece and Whole Word Masking. Whole Word Masking works in a way that if the masked SentencePiece token belongs to a whole word, then all the SentencePiece tokens which form the complete word will be masked altogether. So, only tokens representing entire words are masked, not sub-tokens. The vocabulary size of the Feel-it model is 32005.

[9] https://github.com/musixmatchresearch/umberto

5 Experiments

After reproducing the results presented in (Polignano et al., 2019) on the SENTIPOLC16 dataset, multiple experiments were run in order to compare AlBERTo and Feel-IT, with and without fine-tuning on the target dataset. First, instead of dealing with SENTIPOLC16 by means of 2 independent binary problems (to detect if the tweet has a positive polarity component independently from the negative component) as described in the original AlBERTo's paper, a single classification head with 3 units was used, in order to predict a vector of probability over the three polarity components. We will refer to this model as AlBERTo Multi Class (*AMC*). Second, a minor architectural change involving the removal of the Dropout layer between the [CLS] embeddings and the final classification head, and the changing of training hyper-parameters, was aimed at improving AlBERTo's performances (*AMC opt*) as measured on SENTIPOLC16 for polarity classification. Thus, both AMC and AMC opt inherit the pre-trained AlBERTo's weights, which will be fine-tuned on SENTIPOLC16 for polarity classification, whereas Feel-IT is provided with already fine-tuned weights on the Feel-IT dataset.

A major caveat to be aware of is related to the differences in the output/label domain of both datasets and models. For instance, some datasets are annotated with 3 labels, whereas others with just 2. On the same vein, AMC produces as output 3 labels, whereas Feel-IT was trained to output just 2. This of course leads to some disparities in the evalua-

tion process, which must be taken into account. To alleviate this problem different experiments were conducted, depending on the dataset under analysis. In particular: 1) dataset with 3 output labels (like SENTIPOLC16); run AMC as it is and augment Feel-IT with the output class neutral, which will be predicted for all samples with a binary prediction confidence ≤ 0.65. 2) datasets with 2 output labels (like Feel-IT); run the Feel-IT model as it is and suppress the neutral class prediction of AMC by replacing it with the second most confident prediction. 3) datasets with 3 output labels; since the threshold of 0.65 presented in point 1 for Feel-IT has very often both a very low precision and recall (details available in the SOM), to give a broader overview of its performances remove all neutral samples from the dataset and run another experiment as in point 2 (*no neutrals* experiment).

Then, the fine-tuning experiment involved, for each dataset with a training split, a 5-epochs fine-tuning on the training set, and a testing on the test set. Validation splits are not taken into consideration, since not all datasets have it. AMC and AMC opt undergone a standard fine-tuning (namely *AMC ft* and *AMC opt ft*), while the additional training data was exploited to replace the 2 output units classification head of Feel-IT with a 3 output units head (*Feel-IT ft*), in order to make it more comparable with AMC ft and AMC opt ft.

A final experiment was devoted to the fine-tuning of AMC and Feel-IT on all joined training splits of the datasets. Specifically, AMC was fine-tuned on the concatenation of the training splits of MultiEmotions-IT, Amazon reviews, AriEmozione, Coadapt valence, and TRIP-MAML, whereas Feel-IT was fine-tuned on the training splits of SENTIPOLC16, MultiEmotions-IT, Amazon reviews, AriEmozione, Coadapt valence, and TRIP-MAML. SENTIPOLC16 was left out for AMC since it was already trained on that dataset.

6 Results

Table (5) shows the macro-F1 scores for all pre-trained models, while Table (6) presents the macro-F1 scores for the first fine-tuning experiment. In addition to AlBERTo and Feel-it, baseline F1 scores for a Stochastic Most Frequent Class classifier (SMFC) and an Italian lexicon-based classifier (OpenNER[10]) are provided. Finally, Table (7) shows the macro-F1 scores for the final experiment, in which AMC and Feel-IT are fine-tuned on all training splits jointly.

The detailed results with per-class macro and weighted F1 scores are available in the SOM. As a reference, Table (3) shows the per-class *AMC*'s F1 scores just for the Amazon reviews dataset, from which we can observe an interesting behaviour of the model: the F1 score of the neutral class is extremely low, lowering also the total macro-F1 score for that model. By inspecting some neutral examples in Table (4), we can realize that the Amazon reviews dataset, in particular for the neutral class, contains mislabelled samples, thus causing these low performances. This noise in the annotation was not just observed for the Amazon reviews dataset, but also for other datasets, as can be seen in Table (2) for Trip-maml (see more in the SOM). Noise in the annotations can be addressed by a re-labelling of the datasets or it can be alleviated by using Label Smoothing (Müller et al., 2020) during training.

class	precision	recall	F1	#
negative	0.39	0.75	0.51	75
neutral	0.08	0.11	0.09	38
positive	0.90	0.60	0.72	239

Table 3: Per-class scores and support for AMC on Amazon reviews test

7 Conclusions

This work presented a comparison of 2 BERT-based models on Italian datasets from different domains. Results showed that a BERT-based model trained on tweets can still perform better than a lexicon-based sentiment analyzer (OpenNER in this specific case) for all domains under analysis, and that domain shift for a model trained on tweets seems not that severe, except for the particular case of AriEmozione. Moreover, fine-tuning, even with few thousands samples, is in general very effective in adapting the model to the new domain. However, a deeper analysis is required in order to truly understand how fine-tuning brings new knowledge, and how the model can adapt to a new domain. Further experiments may take into account using larger sequence lengths, grouping together datasets to fine-tune on a specific domain in order to test on another domain, and considering other unrelated domains.

[10] https://www.opener-project.eu/

Text	prediction
Frullatore difficile da montare in più il boccale perde acqua	negative
Buona ! ottimo colore ! proprio come tutti i crazy color sono molto più duraturi degli altri che ho provato forse unica pecca è la tonalità molto chiara infatti consiglio di fare al seguito di una decolorazione per il resto non rovina i capelli quindi consiglio	positive
Quando si attiva é velocissimo ricarica rapidamente ma talvolta fa contatto e la carica non si attiva	neutral

Table 4: Neutral examples of Amazon reviews. Predictions are made by the AMC model

Dataset	SENTIPOLC	Feel-IT	MultiE.	Amazon	Coadapt	Ari	Trip-maml
Domain	Socio-political Tweets	Tweets	YT/FB comments	Product Reviews	Psychology E-therapy	Italian Opera	Hotel Reviews
# Test	1964	2037	447	352	439	250	125
SMFC	0.34 / 0.50*	-	0.37 / 0.49*	0.35 / 0.52*	0.30 / 0.52*	0.30 / 0.45*	0.32 / 0.56*
OpenNER	0.28 / 0.40*	0.59	0.40 / 0.61*	0.35 / 0.56*	0.60 / 0.64*	0.33 / 0.60*	0.50 / 0.74*
AMC	0.64 / 0.76*	**0.89**	**0.66 / 0.82***	0.44 / 0.70*	**0.64 / 0.88***	**0.42** / 0.61*	**0.56** / 0.89*
AMC opt	**0.67** / 0.81*	0.87	0.63 / **0.82***	0.46 / 0.70*	0.57 / 0.84*	0.39 / **0.63***	0.53 / 0.90*
Feel-IT	0.39 / **0.84***	0.99	0.48 / 0.76*	**0.50 / 0.82***	0.31 / 0.81*	0.40 / 0.62*	0.54 / **0.92***

Table 5: Domain, test split size, and F1 scores for all datasets. The experiments marked with * are *no neutral* experiments, as described in Section 5. The Feel-IT model is not chosen as the best result for the Feel-IT dataset since it was trained on it

Dataset	SENTIPOLC	Feel-IT	MultiE.	Amazon	Coadapt	Ari	Trip-maml
Domain	Socio-political Tweets	Tweets	YT/FB comments	Product Reviews	Psychology E-therapy	Italian Opera	Hotel Reviews
# Train	6970	-	2533	820	3417	1962	292
AMC ft	-	-	0.75	**0.62**	**0.78**	0.73	**0.58**
AMC opt ft	-	-	**0.76**	0.52	0.73	0.60	0.57
Feel-IT ft	0.67	-	0.66	0.54	**0.78**	**0.74**	0.57

Table 6: Domain, train split size, and F1 scores for fine-tuned models. Every model is fine-tuned on the training split of the respective dataset, if available. AMC and AMC opt are no further fine-tuned on SENTIPOLC16. In comparison to Table (5) every score increased

Dataset	SENTIPOLC	Feel-IT	MultiE.	Amazon	Coadapt	Ari	Trip-maml
Domain	Socio-political Tweets	Tweets	YT/FB comments	Product Reviews	Psychology E-therapy	Italian Opera	Hotel Reviews
AMC ft	0.63	0.83 ↘	0.75	0.61	0.76 ↘	0.74	0.63 ↗
Feel-IT ft	0.68	0.92 ↘	0.68 ↗	0.66 ↗	0.76 ↘	0.77 ↗	0.73 ↗

Table 7: Domain and F1 scores for fine-tuned models. Every model is fine-tuned on all training splits concatenated together. AMC is not further fine-tuned on SENTIPOLC16. Results are compared with Table (6) and, in case the entry is empty, with Table (5)

References

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. pages 40–45.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22:110–125.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. When does label smoothing help?

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.