Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 581–589
Marseille, 20-25 June 2022

Multi-source Multi-domain Sentiment Analysis with BERT-based Models

Gabriel Roccabruna, Steve Azzolin, Giuseppe Riccardi
Signals and Interactive Systems Lab, DISI, University of Trento
gabriel.roccabruna@unitn.it, steve.azzolin@studenti.unitn.it, giuseppe.riccardi@unitn.it

Abstract

Sentiment analysis is one of the most widely studied tasks in natural language processing. While BERT-based models have achieved state-of-the-art results in this task, little attention has been given to its performance variability across class labels, multi-source and multi-domain corpora. In this paper, we present an improved state-of-the-art and comparatively evaluate BERT-based models for sentiment analysis on Italian corpora. The proposed model is evaluated over eight sentiment analysis corpora from different domains (social media, finance, e-commerce, health, travel) and sources (Twitter, YouTube, Facebook, Amazon, Tripadvisor, Opera and Personal Healthcare Agent) on the prediction of positive, negative and neutral classes. Our findings suggest that BERT-based models are confident in predicting positive and negative examples but not as much with neutral examples. We release the sentiment analysis model as well as a newly financial domain sentiment corpus.

Keywords: sentiment analysis, multi-domain, multi-source

1. Introduction

By now most companies including social media, e-commerce, forecasting companies are collecting and analyzing customers' opinions. These opinions can be extracted from big data produced every day on the Internet using sentiment analysis. Sentiment analysis is the natural language processing task that automatically extracts the writer's orientation from written text (Cardie, 2014). Sentiment analysis has been effectively used to extract attitudes from movie reviews (Kumar et al., 2019), Amazon reviews (Ejaz et al., 2017), and Twitter (Rosenthal et al., 2017).

A widely used technique for sentiment analysis is based on lexicons (Khoo and Johnkhan, 2018; Taboada et al., 2011). Lexicon-based methods assign a score, ranging from positive to negative, to each word in the text. These scores are aggregated into a global score which represents the sentiment polarity associated to a text segment. The most used polarity label dictionary is ternary: *positive, negative, or neutral*. The assumption behind lexicon-based methods is that the sentiment polarity can be inferred from the frequency and the strength of sentiment-charged words, such as `horrible`, `good` or `terrific`. However, a drawback of lexicon-based methods is that they consider the text as a bag of words, limiting the interpretation of natural language in context. The semantic content of a text is a crucial aspect for detecting sentiments that are not anchored to a single word or word chunk. This issue is partially addressed by BERT-based (Kenton and Toutanova, 2019) models, which compute the semantic content representation of a text by exploiting the whole word context.

In recent years, BERT-based models have achieved state-of-the-art results in the sentiment analysis task (Polignano et al., 2019; Xie et al., 2020; Bianchi et al., 2021). However, little research has been conducted on the performance variability of BERT-based models on multi-source and multi-domain corpora. Robustness of BERT-based models would enable these models to be universally applied avoiding the need for fine-tuning process on different kinds of corpora, which is, in some cases, impracticable due to the lack of annotated data. Two important sources of natural language variability are the source and the domain. The source is the channel or website the natural language documents are being extracted from (e.g. Amazon or Twitter), while the domain is the category of the natural language documents such as travel, financial, health. Some sentiment-analysis works on multi-domain corpora refer to reviews of different product categories (Du et al., 2020) and on multi-source corpora refer to reviews coming from different sources, such as Amazon and IMDB (Dai et al., 2020; Khan et al., 2019). However, in both cases the source and domain diversity is low, resulting in a low natural language diversity in terms of jargon, writing style and content. Another issue is that BERT-based models for sentiment analysis are usually evaluated only on positive and negative labels(Polignano et al., 2019; Bianchi et al., 2021), without reporting the performance of the neutral class. However, the neutral class is important since it identifies the absence of a prevalent sentiment.

In this work, we present an improved state-of-the-art BERT-based model for the sentiment analysis task. We have computed the performance variability of our model by comparatively evaluating it on eight corpora in the Italian language of eight different domains and seven different sources on the prediction of positive, negative, and neutral classes. To enrich the domain diversity, we have collected and annotated a novel corpus of financial news and we plan to make it available[1]. The annotation units have been designed to distill the discourse structure relevant for sentiment analysis. We have employed a model to automatically segment the text into functional units, where a functional unit is a concept borrowed from the dialogue act theory (Bunt et al., 2017; Roccabruna et al., 2020). We compare the results of our model with lexicon-based and other neural models. Finally, we conduct an error analysis to identify the most relevant source of errors made by our model. Our findings suggest that BERT-based models are confident in predicting positive and negative classes, but they struggle in predicting the neutral class. This issue is also observed in the inter-annotator agreement, which is lower when neutral examples are considered. Furthermore, our model achieves

[1]The corpus and the sentiment analysis model is available at `http://sisl.disi.unitn.it/itfn-corpus/`

state-of-the-art results on various corpora, showing robust performance across different domains and sources.
The remainder of the paper is organized as follows. Section 2 reviews the related works. In Section 3, we present our model and other models that we used as a comparison. Section 4 describes the corpora used in our experiments. Section 5 presents and compares the evaluations of the tested models. In Section 6, we present and discuss the error analysis. Finally, we present our conclusions.

2. Related works

Sentiment Analysis BERT-based models are effectively used in many natural language processing tasks such as the sentiment analysis task. A common procedure is to start from an off-the-shelf pre-trained model and then fine-tune it on a specific task. However, in many tasks, including sentiment analysis, the fine-tuning needs labeled data, which could be lacking for specific domains such as the health or financial. Nevertheless, little research has been conducted on the performance variability of BERT-based models for sentiment analysis over multi-domain and multi-source corpora to investigate the universal applicability of these models. Moreover, evaluations of BERT-based models for the Italian language report, to the best of our knowledge, only the performance on positive and negative classes, although the class neutral is needed since it identifies the absence of a predominant and clear attitude.
A multi-domain corpus for sentiment analysis for the English language is presented in (Mamta et al., 2020). This corpus is a collection of tweets gathered using a set of keywords, which were selected to identify various social relevant domains. However, the evaluated models are non-BERT-based and the evaluation does not report the performances for each domain. (Du et al., 2020) evaluated BERT-based models on a multi-domain corpus of Amazon reviews, where each domain is a product category. Although the domains used in the evaluation are more than twenty, the domain diversity is low because all the reviews refer to the macro domain of products.
Two multi-source and multi-domain sentiment corpora for the English language were created by (Dai et al., 2020) and (Khan et al., 2019) gathering other existing corpora, which were used to test non-BERT-based models. Although these two corpora differ in size and content, they are composed only of movies and products reviews coming from Amazon and IMDB. Alternatively, (Abid et al., 2020) enhanced the source and domain diversity of their corpus by bringing together existing corpora of tweets, coming from sentiment strength twitter (Thelwall et al., 2012) and Stanford Twitter sentiment Corpus (Go et al., 2017), and movie reviews, coming from IMDB (Maas et al., 2011).
Two of BERT-based publicly available models for the Italian language are AlBERTo (Polignano et al., 2019) and FEEL-IT (Bianchi et al., 2021). AlBERTo is a pre-trained model based on BERT architecture, which was trained on a large number of tweets only on the masked language modelling task. AlBERTo was evaluated on the sentiment analysis task of the SENTIPOLC2016 (Barbieri et al., 2016) corpus. In this evaluation, the authors fine-tuned two separate AlBERTo models, one for predicting the presence of the positive polarity and one for predicting the presence of the negative polarity. FEEL-IT, the other BERT-based model, is a sentiment classifier based on UmBERTo[2], which inherits the RoBERTa architecture and is pre-trained on Commoncrawl ITA. FEEL-IT has only one head used for predicting positive and negative labels. This model was fine-tuned on the homonym corpus, which is a collection of tweets labeled with four non-neutral emotions. However, the performances of FEEL-IT and AlBERTo are not comparable since to train FEEL-IT the authors removed all the neutral examples from the corpus. Nevertheless, there are no explicit evaluations on neutral examples both of AlBERTo and FEEL-IT.

Annotation task Sentiment analysis in the financial field is challenging due to the specific jargon used to express the orientation of news. This specific jargon has to be considered during the definition of the annotation procedure of a sentiment corpus. However, little research has been conducted on the annotation of financial corpora. Indeed, the Financial Phrase Bank (Malo et al., 2014) corpus is one of the few annotated corpora that is thoroughly described and freely available on the web. This corpus is a collection of sentences extracted from financial news articles regarding Finnish companies. The annotation task of this corpus involved the perspective of investors in the annotation of the sentiment polarity. To approximate an actual investor's perspective, the authors of the corpus recruited only financial experts as annotators. The same concept was used by (Takala et al., 2014). However, the need for experts may hinder the annotation of financial sentiment corpora because hiring experts in crowdsourcing is challenging, due to high expenditure and scarce availability.

3. Model description

3.1. Baselines

Prior-Label-Distribution classifier (PLD) is implemented by picking a random class with probability proportional to the distribution of classes in the reference train set.

OpeNER (Open Polarity Enhanced Name Entity Recognition)[3] is an Italian lexicon-based classifier developed as part of a project funded by the European Commission. It can predict the overall polarity of a text, represented as a label among `positive`, `negative` and `neutral`.

3.2. FEEL-IT

The model presented in (Bianchi et al., 2021) uses the Italian model UmBERTo, trained on Commoncrawl ITA for a total of 69GB of raw data, and fine-tuned on the FEEL-IT corpus. UmBERTo inherits from the RoBERTa (Liu et al., 2019) model architecture, which improves the initial BERT by identifying key hyperparameters 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

[2] https://github.com/musixmatchresearch/umberto
[3] https://www.opener-project.eu/

whole word, then all the SentencePiece tokens which form the complete word will be masked altogether. In other words, only tokens representing entire words are masked, not sub-tokens.
FEEL-IT was trained to recognize only positive and negative classes, since the FEEL-IT corpus lacks neutral samples. This leads to a specific evaluation procedure, described in Section 5, to fairly compare the off-the-shelf FEEL-IT with other models.

3.3. AlBERTo Multi-Class

Our first contribution to the AlBERTo architecture (Polignano et al., 2019) consists in using a single multi-class classification head for the labels {`positive, negative, neutral`}, instead of predicting the presence of the positive and negative components separately as presented in the original paper, which requires two independently trained models. To support our decision, we have tested a standard AlBERTo architecture on the multi-class version of the SENTIPOLC16 benchmark for sentiment analysis. Specifically, instead of computing the F1 scores for the two polarity components independently, we have aggregated the two predictions into a single multi-class prediction, then computing the Macro-F1 scores across the three classes. The overall Macro-F1 is 0.64, on par with our multi-class extension obtained by plugging a Softmax activation function over the linearly-projected `[CLS]` embedding:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (1)$$

We will refer to this modified version of AlBERTo as **AMC** (AlBERTo Multi-Class), which inherits the weights of the backbone from the pre-trained AlBERTo published by the authors. The added classification head is learned by fine-tuning the whole network on SENTIPOLC16 with the same hyperparameters used by (Polignano et al., 2019): `weight_decay=0.01`, `learning_rate=`$2e^{-5}$, `num_epochs=3`.
Our additional contribution to AlBERTo is an extensive search of hyperparameters, implemented via the Auto-ML tool Optuna (Akiba et al., 2019). We have run a 50-trials search to optimize `weight_decay`, `learning_rate`, `warmup_ratio` and `num_epochs` validating each epoch on an held-out validation split of SENTIPOLC16. The resulting hyperparameters are then used to train the additional classification head, as described above, over SENTIPOLC16 and to fine-tune the model over different domains for the experiments described in Section 5. Specifically, we have used: `learning_rate=6.599e-05`, `weight_decay=0.0215`, `warmup_ratio=0.899`, `num_epochs=11`. The hyperparameter search can be run for each specific domain, yielding more tailored hyperparameters for the specific domain. However, this would be more prone to overfitting, since the domain-specific corpora are generally small. Furthermore, in the fine-tuning procedure we have used a validation set, which is lacking in (Polignano et al., 2019), by diving in a 90-10 stratified manner the training corpora into a train and a validation split. The model trained with this setting will be henceforth referred to as **AMC opt**. AMC opt outperforms AMC on 6 datasets over 8, as shown in Table 5.

4. Corpora analysis

In this section we present the main resources and the novel financial sentiment corpus for the Italian language. The summary statistics for each corpus are presented in Table 1, whereas examples for each corpus are shown in Table 4. Moreover, in subsection 4.1 we discuss the inter-annotator agreement shown in Table 2.

SENTIPOLC16 (Barbieri et al., 2016) is a collection of socio-politcal Italian Tweets, labelled with subjectivity, sentiment polarity, literal sentiment polarity and irony. The corpus was presented during EVALITA2016 (Barbieri et al., 2016), which is an evaluation campaign of Natural Language Processing and Speech for the Italian language. SENTIPOLC16 is an upgraded version of SENTIPOLC14 (Basile et al., 2021). The main differences with respect to SENTIPOLC14 are the followings. *First*, the test data is composed of a portion of random tweets and a portion of tweets selected via keywords by using a different selection procedure in respect to that used to create the training set. The intention of this was to better assess the generalization capability. *Second*, a portion of the data was annotated via Crowdflower, a crowdsourcing platform, rather than by experts.
The corpus was annotated with `neutral, positive, negative,` and `mixed` classes. The `mixed` class contains examples that are both positive and negative. This class was discarded during our experiments since this class is the least frequent and it is out of the scope of our work.

FEEL-IT (Bianchi et al., 2021) contains Italian Tweets collected between 20th August to 12th October 2020 by monitoring trending topics each day. The corpus is annotated with four emotion labels, which are `joy, sadness, fear` and `anger`. These emotion labels are mapped as {`joy`} $\rightarrow$ `positive` and {`sadness, fear, anger`} $\rightarrow$ `negative`. The annotation was done by two of the authors, who are both Italian native speakers and with Natural Language Processing (NLP) background. The inter-annotation agreement was 0.6 of Krippendorff's Alpha (Krippendorff, 2011). The whole corpus was used to train the FEEL-IT BERT-based model thus, there is no official split in training and test sets. For this reason, in our experiments we have used this corpus as a test set only.

MultiEmotions-IT (MultiE.) (Sprugnoli, 2020) contains manually annotated comments to music videos and advertisements posted on YouTube and Facebook during April 2020. 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. For our experiments, the examples with mixed polarity have been removed from the dataset. The author chose 9 music videos selecting both songs that placed at the top and at the bottom of the popular music contest Sanremo 2020's leaderboard. The annotation was carried out by students with no previous experience in linguistic annotation but with a specific training in the strate-

gic management of communication flows on various media platforms. Overall, each comment was annotated by two students. The inter-annotation agreement was computed using Krippendorff's Alpha (Krippendorff, 2011) and the results are: 0.71 for the positive class, 0.61 for the negative class and 0.38 for the neutral class.

Amazon reviews is a collection of product reviews that we have manually collected via the *Amazon Review Export* Google Chrome extension[4]. We have chosen products of different categories with more than 100 reviews. The reviews have been annotated with the following rules. Reviews with less and more than 3 stars are labeled respectively with negative and positive polarity, whereas reviews with 3 stars are labeled as neutral. However, the lack of a human supervision in the labeling process can bring errors in the annotation, especially for the neutral class, in which we observed various false-neutrals.

Trip-MAML (Jiménez Zafra et al., 2015) was originally intended as a Multi-Aspect Multi-Lingual corpus 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 stars of the reviews in Amazon reviews.

AriEmozione 1.0 (Ari) (Fernicola et al., 2020) 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). The achieved inter-annotator agreement is 0.323 of Cohen's kappa (Cohen, 1960). 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`. The annotators were two Italian native speakers, who annotated all 2,473 instances independently.

Italian Twitter Financial News (ITFN) is a novel financial corpus for the Italian language. To create this corpus, we have collected all the tweets written in three year from 1st January 2018 to 31st December 2020 by four accounts of financial journals, namely Milano Finanza (@MilanoFinanza), Wall Street Italia (@wallstreetita), Italia Oggi (@ItaliaOggi) and Il Sole 24 Ore (@24Finanza). In total, we have collected 84562 tweets. From these, we have annotated 6040 tweets, which have been randomly sampled and stratified on the source. To isolate multiple semantic contents expressed in a text, we have used a BERT-based model trained on dialogue act tagging task to automatically segment the tweets into functional units. In the dialogue act theory, in the ISO standard 24617-2 (Bunt et al., 2017; Roccabruna et al., 2020) functional units are defined as the minimum span of text with a communicative function, which could contain one or more dialogue acts. Although the average length of the collected tweets is only 12 tokens, we have observed that around 1800 tweets are composed of more than one news or different aspects of a news are expressed over multiple functional units. Indeed, the number of functional units per tweet is 1.41 on average, for a total of 8529 functional units.

The annotation task has consisted in identifying the sentiment of a news by looking at the writer's intention. The writer's intention is the way, decided by the writer, in which a news has to be interpreted. This intention can be identified by looking at the style and the terminology used to write the tweet.

Before starting the annotation, the annotators have been trained using our guidelines as reference. In the guidelines, we ask the annotators to consider only the information written in the tweet, without considering any personal knowledge about that company or event, following the hint suggested in (Malo et al., 2014). Moreover, in functional units with ambiguous writer's intended sentiment, the annotators can consider the left and right context in the tweet. Table 3 shows an example of such a case and an example of two functional units being part of the same tweet with opposite polarity.

The annotators have been asked to annotate each functional unit with `positive (1)`, `negative (-1)` or `neutral (0)` valence. Furthermore, we have provided an additional label (`NA`: Not Applicable) to mark spam, non-relevant topics such as gossip or events organized by the journal, and errors made by the model used to segment the tweets. All the functional units annotated with the `NA` label have been removed from the final corpus.

We recruited three annotators; two of them with NLP background and one with psychology background. There was no need for domain (finance) expertise since the tweets are for the general population and the task was to understand the writer's orientation.

The inter-annotator agreement has been computed on 10% of the annotated tweets. The results are in line with other works (Bianchi et al., 2021; Barbieri et al., 2016; Sprugnoli, 2020) and are shown in Table 2. The inter-annotator agreement computed removing all the examples with at least one neutral label shows an improvement from 0.54 to 0.94. The possible motivations of this are discussed in subsection 4.1.

COADAPT valence is a collection[5] of 481 Personal Narratives in the Italian Language (PN) that we have collected by a Personal Healthcare Agent (PHA) in the context of a Digital Cognitive Behavioral Therapy (DCBT) intervention (Mousavi et al., 2021). Each personal narrative has been automatically segmented into functional units by the same model used in ITFN. In this case, the isolation of semantic contents into functional units is even more necessary than in ITFN since the average length of a personal narrative is 65 tokens and they illustrate several events and emotional states lived by users.

The annotation task has been to classify the emotional

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

[5] We are currently applying for further funds to anonymize the corpus and publish a version of the corpus that respects users' privacy and deontological requirements.

valence (Tammewar et al., 2020) of each functional unit with a Likert scale from `-2 (unpleasant)` to `2 (pleasant)`, where `0` is `neutral` valence. In the annotation, three annotators with NLP background have been involved. The inter-annotator agreement score has been computed on 20% of the corpus. The results show a fair agreement and are presented in Table 2. For the task of sentiment analysis, we have collapsed the `negative (-2,-1)`, `positive (1,2)` and `neutral (0)` values in the corresponding classes.

Corpora	#	neg	neutral	pos
SENTIPOLC16	8934	37	42	21
FEEL-IT	2037	64	–	36
Amazon	1172	22	11	67
Multi.E.	2980	26	10	64
Ari	2462	55	2	43
Trip-MAML	417	13	16	71
ITFN	7889	31	35	34
COADAPT	4273	27	60	13

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

Labels	ITFN	COADAPT
-2,-1, 0,1,2	–	0.67
neg, neu, pos	0.54	0.73
neg, pos	0.94	0.98

Table 2: Inter-annotation agreement computed on Italian Twitter Financial News and COADAPT valence. The values are computed with Krippendorff's Alpha (Krippendorff, 2011) using the nominal difference function.

4.1. The Role of Neutrals in the Agreement

Table 2 illustrates the inter-annotator agreement results computed on ITFN and CODAPT. Looking at these results, we can observe that removing all the examples with at least one neutral valence, the inter-annotator agreement score increases of 0.40 points for ITFN and 0.25 points for COADAPT. A problem with the neutral class can be also observed in the MultiEmotions-IT (Sprugnoli, 2020) corpus, in which the neutral class has the lowest agreement compared to the positive and negative classes. This suggests that the neutral class is an important source of disagreement for the annotators. This disagreement might be generated by ambiguities present in the text; we have identified two possible sources of ambiguity. One source dwells in the different meaningful interpretations of some text chunks. The other source is the presence of both positive and negative sentiments in the same text chunk. While for the first source we believe that it is normally present in the natural language, for the second source we have compared the agreement between tweets of ITFN with only one functional unit and tweets with two or more functional units.

ID	Tweets
1	*FU1:* Case - Venezia la città più cara: (**negative**) *Houses - Venice is the most expensive city:* *FU2:* con 200mila euro si compra un appartamento di 45 mq. (*neutral*‖**negative**) *200 thousand euro can buy a flat of 45 square-meter.*
2	*FU1:*Industria, prove di ripresa anche a luglio. (**positive**) *Industry, evidence of recovery also in July.* *FU2:*Ma il crollo tendenziale è dell'8% (**negative**) *But the downward trend is 8%.*

Table 3: FU1 and FU2 are two functional units of a tweet. The first tweet is an example of how functional units influence each other. In this case, FU1 influences with a negative polarity FU2 which would be neutral without a context. The second tweet is an example of two functional units with opposite polarities.

We have observed that the inter-annotator agreement score computed on tweets with just one functional unit is lower (0.51) than the score computed on tweets with more functional units (0.56). Although this analysis is limited by the fact that we have not compared two groups containing the same examples, one with the segmentation into functional units and one without, this analysis brings an evidence to our hypothesis about the source of ambiguity and gives an insight of how this ambiguity can be mitigated by segmenting the text into functional units.

5. Model Evaluation

After reproducing the results presented in (Polignano et al., 2019) and (Bianchi et al., 2021) on the SENTIPOLC16 corpus with the same settings as presented in the paper, multiple experiments have been run in order to compare AMC, AMC opt, and FEEL-IT. The experiments have been organized as follows.

The **first** experiment has been devoted to investigate the performance variability of our proposed AMC opt model across different domains and sources and to compare the performance of AMC opt with other models such as AMC, FEEL-IT, OpeNER, and PLD. In doing this, one issue is the differences in the number of labels both used to annotate the corpora and to train the models. Some corpora are annotated with 3 labels, whereas others with just 2. For the models, the output space of AMC and AMC opt is composed by positive, negative and neutral labels, while FEEL-IT can only output positive and negative labels. This leads to disparities in the evaluation process, which must be taken into account for a fair comparison. To alleviate this problem, we have tested different settings depending on the corpus under analysis. In particular: 1) corpora with 3 output labels (as SENTIPOLC16); we have run all models as they are, except FEEL-IT which was augmented with the output class neutral predicted for all samples with a positive/negative prediction confidence ≤ 0.65. 2) corpora with 2 output labels (as FEEL-IT); we have run the FEEL-IT model as it

Corpus	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
ITFN	La logistica cresce: il Nord-est fa la parte del leone	positive
Amazon	Pessimo acquisto. Durato un mese senza graffi nonostante io abbia applicato la protezione schermo e abbia comprato la cover. Il dispositivo è anche lentissimo e si blocca	negative
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. evitate gente!!	neutral
Ari	Infelice e sventurato potrà farmi ingiusto fato ma infedele io non sarò	positive

Table 4: Examples for each corpus

Corpus	**SENTIPOLC16**	**ITFN**	**FEEL-IT**	**MultiE.**	**Amazon**	**Trip-M.**	**Ari**	**COADAPT**
Source	Twitter			YT/FB	Amazon	Tripadvisor	Opera	PHA
Domain	Socio-political	Financial	General	Comments	Products	Hotels	Opera	Psychology
# Test	1964	785	2037	486	125	352	250	439
PLD	0.34/ 0.50*	0.37/ 0.51*	—	0.37/ 0.49*	0.35/ 0.52*	0.32/ 0.56*	0.30/ 0.45*	0.30/ 0.52*
OpeNER	0.28/ 0.40*	0.40/ 0.59*	0.59	0.40/ 0.61*	0.35/ 0.56*	0.50/ 0.74*	0.33/ 0.60*	0.60/ 0.64*
AMC	0.64/ 0.76*	0.41/ 0.58*	**0.89**	0.66/ 0.82*	0.44/ 0.71*	0.48/ 0.79*	0.42/ 0.61*	0.64/ 0.88*
FEEL-IT	0.39/ 0.84*	0.31/ 0.57*	—	0.48/ 0.76*	0.50/ 0.82*	0.60/ 0.91*	0.40/ 0.62*	0.31/ 0.81*
AMC opt	**0.69/ 0.82***	0.44/ 0.73*	0.87	0.68/ 0.85*	0.51/ 0.78*	0.55/ 0.88*	0.45/ 0.66*	0.64/ 0.82*
AMC opt ss-sd	—	**0.66/ 0.85***	—	0.73/ 0.87*	**0.65/ 0.88***	0.58/ 0.91*	0.73/ 0.74*	0.76/ 0.90*
AMC opt ms-md	0.62/ 0.82*	0.64/ 0.84*	0.87	**0.74/ 0.89***	0.63/ 0.88*	**0.67/ 0.94***	**0.75/ 0.76***	**0.77/ 0.92***

Table 5: Macro-F1 scores for the three types of experiments, along with source, domain, and test split size, for all corpora. The experiments marked with * are *no neutral* experiments, as described in Section 5. The FEEL-IT dataset is used as test set only. SENTIPOLC16 has not been added to the training set of the AMC opt ms-md since AMC opt has already been pre-trained on it.

is and we have suppressed the neutral class prediction of AMC and AMC opt by replacing it with the second most confident prediction. 3) corpora with 3 output labels; since we have observed that, due to overconfidence, the threshold of 0.65 presented in point 1 for FEEL-IT has usually both a very low precision and recall, we have run another variation of the experiment (referred as *no neutral* experiment) in order to compare more faithfully 2 and 3-headed models. Indeed, in this way, all models are tested in both two and three classes settings.

The **ss-sd** experiment (single-source single-domain) has been conceived to assess the maximum performance achievable on each corpus by our AMC opt model, which has scored the highest results for the majority of corpora in the first experiment. To do this, we have fine-tuned the AMC opt model on the training set of each corpus independently. However, the FEEL-IT corpus has not an official split in training and test set. Indeed, in (Bianchi et al., 2021) this corpus has been entirely used either as training or test set. In all our experiments we have used it as a test set. The fine-tuning procedure can be summarized as follows. We have divided the training split of each corpus into 2 sub-splits, one for training and one for validation, in a 90-10 stratified manner with fixed random seed. Early stopping with patience 3 over the held-out validation set has been used to avoid overfitting.

The **ms-md** experiment (multi-source multi-domain) has investigated the possibility of improving the robustness of our model by jointly training the model over a multi-domain and multi-source corpus. The ultimate goal is to verify whether a single model can be universally used to deal with any kind of domain and source effectively. We have followed the simple approach of fine-tuning the models on all joined training splits of the corpora. Specifically, AMC opt has been fine-tuned on the concatenation of the training splits of ITFN, MultiEmotions-IT, Amazon reviews, AriEmozione, COADAPT valence, and TRIP-MAML. SENTIPOLC16 has been left out since the model was pre-trained on that corpus. The fine-tuning procedure is the same as for the ss-sd experiment.
All the corpora used in the experiments have been pre-processed using the pipeline proposed in (Polignano et al., 2019). For ITFN and COADAPT valence corpora, which are annotated at functional unit level, we have performed training and testing considering each functional unit as an independent sample. To avoid possible contamination of

the test set, the corpora have been split into train, validation and test sets at tweet and personal narrative level. We leave to future works the investigation of whether a single BERT-based model can jointly address the segmentation into functional units and sentiment prediction of a text. This could be beneficial since the model could exploit the left and right context of each functional unit as human annotators did.

5.1. Results

Table 5 shows the macro-F1 scores of our experiments. We have grouped the corpus by sources, which are: Twitter, YouTube and Facebook (YT/FB), Amazon, Tripadvisor, Opera and Personal Healthcare Agent (PHA).

Overall, the proposed AMC opt performs on-par or even outperforms the other models in the majority of the corpora, showing that the performance of our model is robust across multi-domain and multi-source corpora.

In addition, the standard single-source single-domain (**ss-sd**) fine-tuning, even with few data, allows the model to achieve better scores on the target domain, resulting in every case in better performances. In some cases, the improvements are huge, such as in ITFN and AriEmozione, in which they have gained respectively 22 and 28 percentage points. In these cases, the jargon and the style of writing are substantially different from the other corpora and, therefore, fine-tuning allows the model to learn the domain peculiarities effectively.

Looking at the **ms-md** and **ss-sd** results, we can observe that the scores achieved by training the model on all corpora (ms-md) are close or even better, such as for Trip-MAML, MultiE. and Ari, to those achieved by fine-tuning the model on single-source and single-domain settings (ss-sd). This shows that our model can be universally used across multi-domain and multi-source corpora attaining good results.

Another aspect that emerged is that the performances of all models are considerably higher when they are assessed only on positive and negative examples. This can be related to the intrinsic simplification of the prediction task, which has one class less, as demonstrated by the increase also in the baseline scores. However, we believe that this highlights an issue with the neutral class, which is further analyzed in Section 6.

6. Error analysis

While Table 5 shows the aggregated results per corpus, a more detailed analysis of the error distribution can shed more light on how the models judges the polarity of a text. In particular, we have analyzed the distribution of F1 scores across corpora for each label, using the AMC opt model as classifier. Figure 1 shows a boxplot with the aforementioned distribution, aggregated by label and divided into two evaluation procedures. The first is obtained by using the standard AMC opt model, while the second one is obtained with a single-source single-domain fine-tuning of AMC opt, as described in Section 5. To compute the boxplots, the results for SENTIPOLC16 and FEEL-IT were excluded. From this plot it is clear that the neutral class is complex. Across different corpora, both the negative and positive polarity components are quite consistently captured, especially after fine-tuning, whereas the neutral class preserves a large variance across corpora and records the lowest performance among the others. We believe that there might be different causes for this. One cause could be that the neutral class is a twofold class, since depending on the annotation guidelines can represent solely a chunk of text with no clearly intended sentiment, or also a chunk of text with a balanced negative-positive contribution. Another cause could be that capturing the neutral component is inherently difficult also for humans. This difficulty adds uncertainty to the neutral class in the corpus, which is subsequently learnt by the model, negatively affecting the performance. As support for this last fact, we report in Table 2 the inter-annotator agreement computed with and without the neutral class and we discuss the possible causes in subsection 4.1.

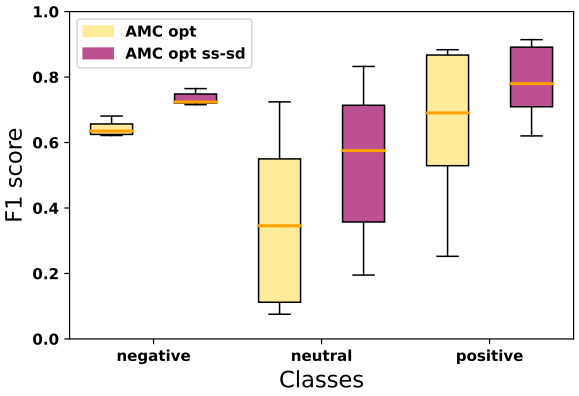

Figure 1: Per-class F1-score distribution across corpora for the AMC opt model and the AMC opt model fine-tuned in a single-source single-domain fashion, (i.e., the fine-tuning is run independently for each corpus).

7. Conclusion

We have presented and evaluated the Italian Twitter Financial News (ITFN) corpus, which is a novel publicly available sentiment corpus in the financial domain for the Italian language. In addition to this, we have introduced AlBERTo Multi-Class (AMC), a modified version of AlBERTo, for predicting negative, positive and neutral classes. Furthermore, with an extensive search of hyperparameters, we have found the best values for AMC obtaining state-of-the-art results for most of the corpora and competitive results for the others. We have named this model AMC optimized (AMC opt) and made it freely available along with ITFN corpus for the research community.

Looking at the results on the performance variability of tested models shown in Table 5, we have observed that our BERT-based model (AMC opt) attains robust performance across different domains and sources. Moreover, we have observed that our model fine-tuned on multi-source and multi-domain corpora jointly achieves good performance compared to the scores achieved fine-tuning the model on the single-source and single-domain setting. This means that the model can learn different aspects coming from different domains and sources and therefore, the model is universally applicable across different domains and sources.

In the error analysis, we have shown that AMC opt, as the other evaluated models, consistently struggles in identifying the neutral class. We have found that this class is problematic also for humans during the annotation process since the inter-annotator agreement is lower when neutral examples are considered. We believe that some possible reasons for this are ambiguities generated by the different meaningful interpretations of some text chunk and the presence of both positive and negative aspects in the same text chunk. While the former is naturally present in the human natural language, for the latter, we have observed that this could be mitigated by the segmentation of the text into functional units. Nonetheless, this needs further research.

8. Acknowledgements

The research leading to these results has received funding from the European Union H2020 Programme under grant agreement 826266: COADAPT.

9. Bibliographical References

Abid, F., Li, C., and Alam, M. (2020). Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks. *Comput. Commun.*, 157:102–115.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016). Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*.

Basile, V., Novielli, N., Croce, D., Barbieri, F., Nissim, M., and Patti, V. (2021). Sentiment polarity classification at evalita: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*, 12:466–478.

Bianchi, F., Nozza, D., and Hovy, D. (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, April. Association for Computational Linguistics.

Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017). Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

Cardie, C. (2014). Sentiment analysis and opinion mining bing liu (university of illinois at chicago) morgan & claypool (synthesis lectures on human language technologies, edited by graeme hirst, 5(1)), 2012, 167 pp; paperbound, isbn 978-1-60845-884-4. *Computational Linguistics*, 40:511–513.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Dai, Y., Liu, J., Ren, X., and Xu, Z. (2020). Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *AAAI*.

Du, C., Sun, H., Wang, J., Qi, Q., and Liao, J. (2020). Adversarial and domain-aware bert for cross-domain sentiment analysis. In *ACL*.

Ejaz, A., Turabee, Z., Rahim, M., and Khoja, S. A. (2017). Opinion mining approaches on amazon product reviews: A comparative study. *2017 International Conference on Information and Communication Technologies (ICICT)*, pages 173–179.

Fernicola, F., Zhang, S., Garcea, F., Bonora, P., and Barrón-Cedeño, A. (2020). Ariemozione: Identifying emotions in opera verses. In Johanna Monti, et al., editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Go, A., Bhayani, R., and Huang, L. (2017). For academics-sentiment 140-a twitter sentiment analysis tool.

Jiménez Zafra, S. M., Berardi, G., Esuli, A., Marcheggiani, D., Martín-Valdivia, M. T., and Moreo Fernández, A. (2015). A multi-lingual annotated dataset for aspect-oriented opinion mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2533–2538, Lisbon, Portugal, September. Association for Computational Linguistics.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Khan, F. H., Qamar, U., and Bashir, S. (2019). Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach. *Soft Computing*, 23:5431–5442.

Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Kumar, H. M. K., Harish, B. S., and Darshan, H. K. (2019). Sentiment analysis on imdb movie reviews using hybrid feature extraction method. *Int. J. Interact. Multim. Artif. Intell.*, 5:109–114.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *ACL*.

Malo, P., Sinha, A., Korhonen, P. J., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Mamta, Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet cor-

pora for sentiment analysis: Resource creation and evaluation. In *LREC*.

Mousavi, S. M., Cervone, A., Danieli, M., and Riccardi, G. (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, June. Association for Computational Linguistics.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (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.

Roccabruna, G., Cervone, A., and Riccardi, G. (2020). Multifunctional iso standard dialogue act tagging in italian. In *CLiC-it*.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In **SEMEVAL*.

Sprugnoli, R. (2020). Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian. 12.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 06.

Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*.

Tammewar, A., Cervone, A., Messner, E.-M., and Riccardi, G. (2020). Annotation of emotion carriers in personal narratives. In *LREC*.

Thelwall, M. A., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *J. Assoc. Inf. Sci. Technol.*, 63:163–173.

Xie, Q., Dai, Z., Hovy, E. H., Luong, M.-T., and Le, Q. V. (2020). Unsupervised data augmentation for consistency training. *arXiv: Learning*.