ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Antici, F., Bolognini, L., Inajetovic, M.A., Ivasiuk, B., Galassi, A., Ruggeri, F. (2021). SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers [10.1007/978-3-030-85251-1_4].

Availability:
This version is available at: https://hdl.handle.net/11585/832327 since: 2021-10-05

Published:
DOI: http://doi.org/10.1007/978-3-030-85251-1_4

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

15 October 2026

This is the final peer-reviewed accepted manuscript of:

Antici, F., Bolognini, L., Inajetovic, M.A., Ivasiuk, B., Galassi, A., Ruggeri, F. (2021). SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers. In: , et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, vol 12880. Springer, Cham, pp. 40–52

The final published version is available online at ***https://dx.doi.org/10.1007/978-3-030-85251-1_4***

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Paper accepted at CLEF 2021, International Conference of the Cross-Language Evaluation Forum for European Languages, Online Event, September 21–24, 2021.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-85251-1_4.

Please cite this work as:
Antici F., Bolognini L., Inajetovic M.A., Ivasiuk B., Galassi A., Ruggeri F. (2021) SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers. In: Candan K.S. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, vol 12880. Springer, Cham. https://doi.org/10.1007/978-3-030-85251-1_4

SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers

Francesco Antici*, Luca Bolognini*, Matteo Antonio Inajetovic*, Bogdan Ivasiuk*, Andrea Galassi(✉)[0000−0001−9711−7042], and Federico Ruggeri(✉)[0000−0002−1697−8586]

DISI, University of Bologna, Bologna, Italy
{francesco.antici,luca.bolognini3}@studio.unibo.it
{matteo.inajetovic,bogdan.ivasiuk}@studio.unibo.it
{a.galassi,federico.ruggeri6}@unibo.it

Abstract. We present SubjectivITA: the first Italian corpus for subjectivity detection on news articles, with annotations at sentence and document level. Our corpus consists of 103 articles extracted from online newspapers, amounting to 1,841 sentences. We also define baselines for sentence- and document-level subjectivity detection using transformer-based and statistical classifiers. Our results suggest that sentence-level subjectivity annotations may often be sufficient to classify the whole document.

Keywords: Subjectivity Detection · Italian Language · News Articles · Natural Language Processing · Deep Learning

1 Introduction

Subjectivity detection (SD) consists of understanding whether a given piece of text is biased by its creator or not. As highlighted by Chaturvedi et al. [6], SD is a very complex task because the perception of subjectivity is subjective in itself and may derive from different levels of expertise, different interpretations of the language, and also conscious and unconscious biases linked to the personal background. Moreover, domains characterized by the lack of context, such as Tweets, or by references and quotes, such as news articles, pose an additional challenge.

The ability to detect subjectivity in textual documents can greatly help other tasks [31] such as fake news detection, information extraction, question answering, sentiment analysis, and argument mining. The recent success of machine learning techniques based on deep neural networks in many NLP tasks has partially relieved the need for structured knowledge, but it has increased the need for labeled corpora for training. While many resources exist for the English language, the same can not be said for other ones. Projection techniques [12,17]

* Equal contribution.

Table 1. Nonexhaustive list of SD and SSA corpora. The Size column refers to the type of elements in the Granularity column.

Datset	Task	Domain	Language	Granularity	Size
Wiebe et al. [30]	SD	News	English	Sentence	$\sim$ 500
Chesley et al. [8]	SSA	News	English	Document	$\sim$ 1,000
Movie Review [21]	SSA	Reviews	English	Document	$\sim$ 2,000
MOAT [7,26]	SSA	News	English	Sentence	$\sim$ 3,500
MPQA [7,32]	SSA	News	English	Sentence	$\sim$ 16,000
NoReC$_{fine}$ [20]	SA	News	Norwegian	Sentence	$\sim$ 8,000
MSA [1]	SSA	News	Arabic	Sentence	$\sim$ 3,000
Odia [18]	SSA	News	Odia	Sentence	$\sim$ 2,000
Volkova et al. [29]	SSA	Twitter	Eng, Spa, Rus	Tweet	$\sim$ 4,500,000
Senti-TUT [4]	SSA	Twitter	Italian	Tweet	$\sim$ 3,000
Felicitta [3]	SSA	Twitter	Italian	Tweet	$\sim$ 1,000
SubjectivITA	SD	News	Italian	**Sent.+Doc.**	1,841 S; 103 D

can be used to create new corpora in an unsupervised fashion, but they usually need parallel corpora or they rely on automatic translation processes that may compromise the subjective form that some words have in the original language. Additionally, the lack of non-English corpora hinders the evaluation of any cross-lingual technique.

For these reasons, we have created SubjectivITA, the first corpus for SD made of newspaper articles in the Italian language. The corpus has been manually annotated at two different levels of granularity, therefore it is suitable to perform the task both at sentence and document level. To guarantee the quality of the corpus, we followed an iterative process of discussion and modification of the guidelines, so as to align the opinions of the annotators and increase their agreement. We report the problems that emerged during this process and discuss how to address specific ambiguous cases. Finally, we used our corpus as a benchmark to evaluate a set of machine learning techniques that range from basic methods such as Logistic Regression, to state of the art NLP models like BERT [11].

In Section 2 we survey related works and present a comparison between existing corpora. In Sections 3 we describe our labeling process and our guidelines. In Section 4 we present our experimental evaluation, while in Section 5 we draw conclusions and discuss possible future developments.

2 Related Work

SD is a well-known task, and over the years many resources and methods to address it have been developed. We focus our attention on existing corpora, in particular on those that address this task specifically and those that are not in the English language, framing them in Table 1 according to multiple aspects. A more comprehensive overview of the topic, including the evolution of

SD over the years and the relation with other tasks, is covered by the excellent survey of Chaturvedi et al. [6]. To the best of our knowledge, Wiebe et al. [30] are the first to create a corpus for SD. They annotate a set of news articles and also describe an iterative process to improve inter-annotator agreement and annotation guidelines, from which we draw inspiration for our own process.

Since subjective sentences and documents usually express a stance towards a topic, SD and sentiment analysis (SA) can be performed together. Chesley et al. [8] present one of the first Subjectivity and Sentiment Analysis (SSA) corpora, a multi-class classification task where labels specify whether a subjective sentence conveys a positive or negative sentiment. In recent SSA corpora based on news sources in non-English languages [1,18,20], documents are tagged at the sentence level to obtain more fine-grained labeling than the one achievable by using only document-level tagging. In our corpus, documents are labeled both at the sentence and document level, but we use only two labels (*Subjective* and *Objective*). Concerning the Italian language, existing SSA corpora are mainly based on Twitter [3,4], while our proposed corpus has been obtained from Italian newspapers. Our annotation process is very similar to the one described by Bosco et al. [3], except for some differences in the partition of tasks assigned to the annotators.

3 Creation of the Corpus

Our SubjectivITA corpus was created by manually gathering articles from Italian online newspapers, chosen so as to cover a wide spectrum of styles and topics. The choice fell on outlets of national importance and usually considered as politically impartial, but also on local outlets, columns, and blogs, hoping thus to include more subjective content. The articles were collected between the 20th of January 2021 and the 1st of February 2021 and were chosen randomly among those that contained less than 40 sentences. Both the corpus and the guidelines (in the Italian language) are publicly available.[1]

3.1 Annotation Process

The articles were annotated using two labels, *Objective* (**OBJ**) and *Subjective* (**SUBJ**),[2] defined in Section 3.2. Following an initial guidelines draft, four Italian native speakers (A1, A2, A3, A4) independently annotated the same set of 6 articles, totalling 80 sentences, obtaining a preliminary small corpus named $\mathbf{P_1^I}$. The annotation phase consisted of the following step:

1. *Segmentation*: the articles are manually split into separate sentences.[3]

[1] https://github.com/francescoantici/SubjectivITA

[2] The original Italian terms and labels are "OGGettivo" and "SOGgettivo".

[3] Since different authors have different styles of writing and follow different conventions regarding punctuation symbols, we preferred to not rely on automatic segmentation tools since they may introduce errors.

Table 2. Cohen's kappa results on **sentences** tags.

(a) Corpus $\mathbf{P_1^I}$

	A1	**A2**	**A3**	**A4**
A1	-	0.38	0.21	0.44
A2	0.38	-	0.41	0.36
A3	0.21	0.41	-	0.51
A4	0.44	0.36	0.51	-

(b) Corpus $\mathbf{P_1^F}$

	A1	**A2**	**A3**	**A4**
A1	-	0.52	0.59	0.52
A2	0.52	-	0.66	0.82
A3	0.59	0.66	-	0.73
A4	0.52	0.82	0.73	-

(c) Corpus $\mathbf{P_2}$

	A1	**A2**	**A3**	**A4**
A1	-	0.52	0.50	0.51
A2	0.52	-	0.65	0.66
A3	0.50	0.65	-	0.76
A4	0.51	0.66	0.76	-

Table 3. Cohen's kappa results on **articles** tags.

(a) Corpus $\mathbf{P_1^I}$

	A1	**A2**	**A3**	**A4**
A1	-	0.67	0.33	0.67
A2	0.67	-	0.67	0.25
A3	0.33	0.67	-	0.00
A4	0.67	0.25	0.00	-

(b) Corpus $\mathbf{P_1^F}$

	A1	**A2**	**A3**	**A4**
A1	-	0.25	0.57	0.25
A2	0.25	-	0.57	1.00
A3	0.57	0.57	-	0.57
A4	0.25	1.00	0.57	-

(c) Corpus $\mathbf{P_2}$

	A1	**A2**	**A3**	**A4**
A1	-	0.53	0.37	0.53
A2	0.53	-	0.78	0.55
A3	0.37	0.78	-	0.78
A4	0.53	0.55	0.78	-

2. *Sentence Labeling*: each sentence obtained from step 1 is labeled independently of the context (i.e. the other sentences).
3. *Document Labeling*: after all the sentences have been labeled, the article is evaluated in its entirety and the appropriate label is assigned.

Then, a guideline improvement phase followed, achieved through group discussion and the annotators' feedback. In this phase, guidelines were refined and expanded to cover unforeseen situations and clarify the ambiguities on which the annotators were either doubtful or disagreeing on. Such a process on annotation and guidelines improvement was iterated multiple times, monitoring the annotators' agreement, until the quality of the annotations was considered satisfactory.

The agreement between the annotators was measured using *Cohen's Kappa* and *Fleiss' Kappa*, and it is shown in Tables 2, 3 and 4. The agreement between each pair of annotators was assessed through Cohen's Kappa, to study the correlation between the annotators, monitor interpretation biases, and make the evaluation transparent. For example, Table 3a clearly shows that after the first iteration, annotators A3 and A4 had no agreement on document annotation. Fleiss' kappa was instead used to monitor the agreement of the whole group, and was used as the stopping criterion: the iterative process finished once substantial agreement ($\kappa \geq 0.6$) [16] on the sentence-level annotation was reached.

This process significantly improved the agreement between the annotators, as clearly shown in Table 4, and led to the final version of this preliminary corpus, named $\mathbf{P_1^F}$. Once the the guidelines were finished, they were validated by creating a new preliminary corpus **P2** and evaluating the agreement between the annotators. Such a corpus was composed of 9 articles, amounting to 145

Table 4. Fleiss' kappa values on tags.

Level	P_1^I	P_1^F	P_2
Sentence	0.24	0.65	0.61
Article	0.30	0.53	0.58

Table 5. Summary of the guidelines for sentence tagging.

Objective Rules	Subjective Rules
O1) Report, historic events, or statistics	S1) Explicit personal opinion
O2) Report of a third subject's emotions	S2) Ironic or sarcastic expression
O3) No conclusions without supporting data	S3) Personal wishes and hopes
O4) Conclusions supported by data	S4) Discriminating expressions
O5) Public and commonly used nicknames	S5) Exaggerated expressions
O6) Common sayings	S6) Conclusions not supported by data
O7) Absence of explicit personal opinions	S7) Expression of subjective emotion
O8) No other rule applies	

sentences, on which the agreement between the annotators both at sentence- and article-level was close to substantial. The guidelines were therefore considered a reliable tool and were used to annotate the remaining articles. Each annotator received 22 different articles, which they tagged individually, resulting in a final corpus of 103 articles with a total of 1,841 sentences.

3.2 Definition of Objective and Subjective

We define a sentence as subjective whenever it shows its author's point of view or opinion on the matter, even if it's only using irony or sarcasm. Otherwise, the sentence is considered objective. The same definition applies when labeling documents: they are considered subjective when they express, to some degree, the author's personal opinions on the topic at hand, and objective otherwise. The labelling of the documents must not rely on a quantitative evaluation of the number of objective and subjective sentences, but instead on the characteristics of the document as a whole. These general definitions have been further developed in the guidelines as a set of specific rules that have been used by the annotators to discriminate ambiguous cases. We list these rules in Table 5, while in Table 6 we report examples of sentences from the corpus and the guidelines, specifying which rule was applied to label them.

It is important to underline some aspects related to these rules and our decisions regarding ambiguous cases. First of all, since the context of the sentences is not considered for their annotation, sentences that are objective by themselves are labeled as such, even if they would be considered subjective in the specific context of the article where they belong. This may be the case, for example, of sentences that contain subtle irony. Moreover, any fact or data reported in

Table 6. Examples of sentences (translated from Italian) with their respective tag and the annotation rule that was applied.

Sentence	Tag	Rule
Without school, Andrew's day is never ending.	OBJ	O2
I hope Renzi sues him.	SUBJ	S3
They celebrated as if there was no coronavirus.	SUBJ	S1
28 December 1977: the New Partisans kill Angelo Pistolesi.	OBJ	O1
The consumer expressed his disappointment in a web post.	OBJ	O2
You are the worst administration.	SUBJ	S4
Supplies seems to be available at international level, but it isn't clear yet.	OBJ	O3

the articles is assumed to be true, unless they concern something that is widely known as incorrect (e.g., *The Sun revolves around the Earth*).

One of the most controversial cases of discussion is how quotes influence the subjectivity of an article. Quotes in news articles usually report the words of a person that expresses their personal and subjective perspective on a topic. Since sentences are annotated independently of the context, those that contain quotes are likely to be classified as subjective. In cases where a journalist addresses a topic without expressing their own perspective and only reporting other people's opinions, we can say that the article, in its entirety, is objective. That will therefore result in a document that contains mostly subjective sentences, but it is objective in itself. It can be argued that the best practice to address a controversial topic is to report quotes from parties with different opinions. However, when only one of those parties' opinion is considered by the author, neglecting the others, then the article may aim to influence the reader and skews the perspective towards being subjective. We have chosen to not address this specific case due to its complexity, and to leave it as subject for future work.

Another ambiguous case is whether hypotheses brought up by the author without supporting data should be considered subjective. We decided to distinguish two cases. If the author proposes a hypothetical development of the considered matter and presents it as the only possible scenario, the sentence is considered subjective (rule S6). Conversely, if the development is proposed just as a possible interpretation yet no accent is placed on the veracity of this hypothesis, then the sentence is labeled as objective (rule O3). Obviously, in cases where hypotheses are directly supported by reported facts and data, the sentence is considered objective (rule O4).

4 Subjectivity Detection

Subjectivity detection can be tackled at different levels of granularity depending on the considered textual units that have to be classified. In our experimental setup, we explore the tasks of sentence- and document-level subjectivity detection. We formulate both tasks as a binary classification problem where an input example x can either be subjective or objective.

Table 7. Classification performance for sentence-level subjectivity detection. We report precision, recall and F1-score for the subjective class **SUBJ**. Additionally, we also consider summary metrics like accuracy and F1-macro scores.

Model	P-SUBJ	R-SUBJ	F1-SUBJ	Accuracy	F1-macro
GRU	0.46	**0.73**	0.56	0.63	0.62
MulilingualBERT	**0.62**	0.67	**0.64**	**0.76**	**0.73**
AlBERTo	**0.62**	0.65	0.63	0.75	0.72
MAJ-B	0.0	0.0	0.0	0.67	0.40
WR-B	0.33	0.30	0.32	0.57	0.50

Table 8. SubjectivITA corpus statistics for subjectivity detection.

(a) Dataset statistics for sentence-level SD.

Split	SUBJ	OBJ	Total
Train	401	998	1,399
Validation	81	134	215
Test	75	152	227

(b) Dataset statistics for document-level SD.

Split	SUBJ	OBJ	Total
Train	28	46	74
Test	10	19	29

In particular, document-level classification is a task that comes with multiple valid formulations, where the simplest of them consists in aggregating sentence-level predictions into a single result. Certainly, subjective sentences may have an impact on the overall document label, but when we increase the scope to whole documents, we have also to consider other relevant factors, such as each sentence context, relations, and overall contribution to the gist of the document itself. For instance, a document may contain some subjective sentences that have a marginal contribution to its narrative point of view, thus, not sufficiently impacting the discourse to alter the perceived perspective. Conversely, a document that contains mostly objective sentences may end with a very subjective conclusion, shifting towards subjectivity. Nonetheless, solely focusing on sentence-level subjectivity annotations still represents a valuable baseline worth considering.

4.1 Sentence-level Detection

Problem Description. In the context of sentence-level subjectivity detection, an input x is represented by a sentence contained in our corpus. Our approach follows an end-to-end perspective by considering deep learning models that directly encode x via an embedding layer and assign it a label $\tilde{y} \in \{\textbf{SUBJ}, \textbf{OBJ}\}$.

Models. We consider two major classes of deep learning models in our experimental setup: a) recurrent neural networks and b) transformer-based architectures. Due to the unbalance of our corpus, we also consider a majority baseline, namely **MAJ-B**, and a weighted random baseline based on class distribution, **WR-B**. The models we evaluate are the following:

– **Bi-GRU**: a single-level bi-directional GRU [9] followed by a single dense layer for classification. The employed configuration is as follows: 16 units for the GRU layer with 0.1 dropout rate and 1 unit with sigmoid activation for the dense layer. We consider pre-trained GloVe [22] with embedding dimension set to 200.
– **MulilingualBERT**: the pre-trained `bert-base-multilingual-uncased` version of BERT.[4] As in most of NLP task, fine-tuned BERT [11] models have been successfully used to address SD [14] and related NLP tasks [10,15,19].
– **AlBERTo** [23]: a pre-trained version of BERT for the Italian language, initially fine-tuned on Italian tweets for the task of sentiment analysis[4]. We consider this model due to the success of BERT-based models on Italian language tasks [27].

Methodology. We divided the corpus sentences into three splits, by randomly assigning the documents to train (75%), validation (12.5%), and test set (12.5%). Table 8a reports a summary of the dataset composition. As a preliminary step, we carried out a hyper-parameter calibration routine by picking the best configuration based on the performance achieved on the validation set. Given the small amount of available data and the non-deterministic aspect of neural networks [24], we repeatedly trained each neural model on the train set with different random seed initialization. We set the number of repetitions to 3. We regularized by early stopping the training phase based on the validation accuracy score. Concerning model optimization, each model was trained to minimize a binary cross-entropy loss. The **Bi-GRU** baseline had the learning rate set to 0.01 and uses Adam optimizer. Both **BERT** and **AlBERTo** had their learning rate set to 1e-5 after the calibration phase. All models were trained for a maximum of 30 epochs and had the early stopping patience is set to 3.

Results. Table 7 summarizes the results of the sentence-level subjectivity detection task. Each metrics is to be considered as the average over three individual model runs. All employed deep learning models are well above the majority baseline **MAJ-B**. In particular, the **GRU** baseline reaches satisfactory performance with a 0.62 F1-macro score and 0.56 F1-SUBJ score, but it is significantly outperformed by the BERT-based models. **MulilingualBERT** and **AlBERTo** achieve comparable performance, with the former achieving few percentage points more. Due to the challenging nature of the task and the imperfect agreement between annotators, it is difficult to evaluate what is the upper bound on this task and how much space for improvement there is.

4.2 Document-level Detection

Problem Description. In the sentence-level setting, the inputs x are represented by the documents of our corpus. In this scenario, we opt for a more

[4] For all the transformer architectures we considered the implementations available at http://huggingface.co/.

Table 9. Classification performance for document-level subjectivity detection. We mainly report precision, recall, and F1-score for the subjective class **SUBJ**. Additionally, we also consider summary metrics like accuracy and F1-macro scores.

Model	P-SUBJ	R-SUBJ	F1-SUBJ	Acc	F1-macro
RF	0.56	0.50	0.53	0.69	0.65
DT	0.60	0.30	0.40	0.69	0.60
SVM	**0.83**	0.50	0.62	**0.79**	**0.74**
LR	0.80	0.40	0.53	0.76	0.69
MAJ-B	0.0	0.0	0.0	0.66	0.40
WR-B	0.39	0.43	0.41	0.57	0.54
r-SUBJ	0.77	**0.89**	**0.83**	0.76	0.71

traditional machine learning approach, mainly due to the small corpus size. As already stated, document-level detection cannot be solely reduced to a function of sentence-level predictions because it involves multiple factors like contextual information and relevance to the document narrative. For this reason we hypothesize that deep learning approaches applied to the whole document would probably lead to better results. Nonetheless, in this stage of work we are mainly interested in presenting valuable baselines for the task. In particular, we evaluate to what extent features that mainly concern sentence-level subjectivity labels can be considered reliable. On this basis, we manually select a set of hand-crafted features that sums up the content of each article concerning subjectivity information. More precisely, we consider for each article the following indicators: number of sentences, number of objective sentences, number of subjective sentences, and article source.

Models. We consider the following set of linear classifiers: Random Forest (**RF**), Decision Tree (**DT**), Support Vector Machine (**SVM**), and Logistic Regression (**LR**).[5] We consider the same baselines described for sentence-level detection. In addition, we consider a threshold-based baseline, namely **r-SUBJ**, which discriminates between subjective and objective articles based on the average ratio of subjective sentences per article computed on the train set.

Methodology. We initially randomly split collected news articles into train (70%) and test (30%) sets, respectively, obtaining the dataset illustrated by Table 8b. Models are initially trained on the train set and later evaluated on the test set. No preliminary hyper-parameter calibration phase was considered in this scenario.

[5] All mentioned models are employed with their default configuration as defined within the `sciki-learn` python library: http://scikit-learn.org/stable/.

Results. Table 9 summarizes obtained results for each model. In particular, **SVM** significantly outperforms other machine learning models, achieving an F1-SUBJ and F1-macro scores of respectively 0.62 and 0.74. Surprisingly, the ratio-based baseline **r-SUBJ** achieves the highest F1-SUBJ score (0.83) and is second only to **SVM**. Such results favor the simplifying hypothesis that even summary sentence-level subjectivity information is a useful indicator for this task. Overall, all reported models would certainly benefit from a preliminary hyper-parameters calibration phase.

5 Conclusions

We presented a new Italian corpus for subjectivity detection in news articles. To the best of our knowledge, this is the first Italian corpus language to address this domain and also to have annotations both at document and sentence level. During the annotation we have encountered and discussed problems related to the inherent ambiguity of the task at hand, such as sentences involving quotes and irony, resulting in the creation of detailed guidelines that may help the creation new future resources. Finally, we produced a few baselines. Our results suggest that sentence-level information may be enough to properly classify documents, even if it may lead to misclassification of some ambiguous cases, such as documents with many quotes. We plan to test this hypothesis in future works.

Due to the scarcity of similar resources, the corpus is meant to contribute to research in SD, but it could also be used in multi-objective or transfer learning settings. This could be done across different dimensions, such as domains (news and tweets), languages, and related tasks (e.g., sentiment analysis [5], argument mining [2], and fake news detection [28]). Future research directions include extending the corpus, allowing a better and more robust evaluation of deep learning solutions and enriching the corpus with additional annotation layers concerning strongly correlated tasks like sentiment analysis. A further possibility would be to operate with a non-binary subjectivity scale in the hope that a richer annotation scheme might improve the effectiveness of SD as an auxiliary task. However, the definition of such a scale would pose additional challenges. For what concerns the experimental part, we aim to apply more advanced techniques to the document-level detection, exploiting sentence embeddings [25] and hierarchical architectures based on neural attention [13,33]. Finally, we plan to perform experiments regarding transfer learning across corpora in different languages exploiting automatic translation of documents.

Acknowledgement

We would like to thank Paolo Torroni for his help and supervision.

References

1. Abdul-Mageed, M., Diab, M.T.: Subjectivity and sentiment annotation of modern standard arabic newswire. In: Linguistic Annotation Workshop LAW. pp. 110–118. The Association for Computer Linguistics (2011)
2. Basile, P., Basile, V., Cabrio, E., Villata, S.: Argument mining on italian news blogs. In: CLiC-it/EVALITA. vol. 1749. CEUR-WS.org (2016)
3. Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G.F., Sanguinetti, M., Sulis, E.: Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicitta. In: ES^{3}LOD@LREC. pp. 56–63. ELRA (2014)
4. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis: The case of irony and senti-tut. IEEE Intell. Syst. **28**(2), 55–63 (2013). https://doi.org/10.1109/MIS.2013.28
5. Caselli, T., Novielli, N., Patti, V., Rosso, P.: Evalita 2018: Overview on the 6th evaluation campaign of natural language processing and speech tools for italian. In: EVALITA@CLiC-it. vol. 2263. CEUR-WS.org (2018)
6. Chaturvedi, I., Cambria, E., Welsch, R.E., Herrera, F.: Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. Information Fusion **44**, 65–77 (2018). https://doi.org/10.1016/j.inffus.2017.12.006
7. Chenlo, J.M., Losada, D.E.: An empirical study of sentence features for subjectivity and polarity classification. Inf. Sci. **280**, 275–288 (2014). https://doi.org/10.1016/j.ins.2014.05.009
8. Chesley, P., Vincent, B., Xu, L., Srihari, R.K.: Using verbs and adjectives to automatically classify blog sentiment. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 27–29. AAAI (2006)
9. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP. pp. 1724–1734. ACL (2014). https://doi.org/10.3115/v1/d14-1179
10. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A.S., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. In: ACL. pp. 4040–4054. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.372
11. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1). pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/n19-1423
12. Galassi, A., Drazewski, K., Lippi, M., Torroni, P.: Cross-lingual annotation projection in legal texts. In: COLING. pp. 915–926. International Committee on Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.coling-main.79
13. Galassi, A., Lippi, M., Torroni, P.: Attention in natural language processing. IEEE Transactions on Neural Networks and Learning Systems pp. 1–18 (2020). https://doi.org/10.1109/TNNLS.2020.3019893
14. Huo, H., Iwaihara, M.: Utilizing BERT pretrained models with various fine-tune methods for subjectivity detection. In: APWeb/WAIM (2). vol. 12318, pp. 270–284. Springer (2020). https://doi.org/10.1007/978-3-030-60290-1_21
15. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia Tools and Applications pp. 1–24 (2021). https://doi.org/10.1007/s11042-020-10183-2
16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977). https://doi.org/10.2307/2529310

17. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: ACL. ACL (2007)
18. Mohanty, G., Mishra, P., Mamidi, R.: Annotated corpus for sentiment analysis in odia language. In: LREC. pp. 2788–2795. ELRA (2020)
19. Mozafari, M., Farahbakhsh, R., Crespi, N.: A bert-based transfer learning approach for hate speech detection in online social media. In: COMPLEX NETWORKS (1). vol. 881, pp. 928–940. Springer (2019). https://doi.org/10.1007/978-3-030-36687-2_77
20. Øvrelid, L., Mæhlum, P., Barnes, J., Velldal, E.: A fine-grained sentiment dataset for norwegian. In: LREC. pp. 5025–5033. European Language Resources Association (2020)
21. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: ACL. pp. 271–278. ACL (2004). https://doi.org/10.3115/1218955.1218990
22. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543. ACL (2014). https://doi.org/10.3115/v1/d14-1162
23. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., Basile, V.: Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: CLiC-it. vol. 2481. CEUR-WS.org (2019)
24. Reimers, N., Gurevych, I.: Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In: EMNLP. pp. 338–348. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/d17-1035
25. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: EMNLP (1). pp. 4512–4525. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.365
26. Seki, Y., Evans, D.K., Ku, L., Sun, L., Chen, H., Kando, N.: Overview of multilingual opinion analysis task at NTCIR-7. In: NTCIR. National Institute of Informatics (NII) (2008)
27. Tamburini, F.: How "BERTology" changed the state-of-the-art also for italian NLP. In: CLiC-it. vol. 2769. CEUR-WS.org (2020)
28. Vedova, M.L.D., Tacchini, E., Moret, S., Ballarin, G., Pierro, M.D., de Alfaro, L.: Automatic online fake news detection combining content and social signals. In: FRUCT. pp. 272–279. IEEE (2018). https://doi.org/10.23919/FRUCT.2018.8468301
29. Volkova, S., Wilson, T., Yarowsky, D.: Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: EMNLP. pp. 1815–1827. ACL (2013)
30. Wiebe, J., Bruce, R.F., O'Hara, T.P.: Development and use of a gold-standard data set for subjectivity classifications. In: ACL. pp. 246–253. ACL (1999)
31. Wiebe, J., Wilson, T., Bruce, R.F., Bell, M., Martin, M.: Learning subjective language. Comput. Linguistics **30**(3), 277–308 (2004)
32. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Lang. Resour. Evaluation **39**(2-3), 165–210 (2005). https://doi.org/10.1007/s10579-005-7880-9
33. Zhang, X., Wei, F., Zhou, M.: HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In: ACL (1). pp. 5059–5069. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1499