A Machine Learning approach for Sentiment Analysis for Italian Reviews in Healthcare

Luca Bacco[1,2,3], Andrea Cimino[2], Luca Paulon[3], Mario Merone [1], Felice Dell'Orletta[2]
[1]Università Campus Bio-Medico (UCBM)
[2] Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR), ItaliaNLP Lab
[3]Webmonks s.r.l.
{l.bacco, m.merone}@unicampus.it
paulon@webmonks.it
{andrea.cimino, felice.dellorletta}@ilc.cnr.it

Abstract

In this paper, we present our approach to the task of binary sentiment classification for Italian reviews in healthcare domain. We first collected a new dataset for such domain. Then, we compared the results obtained by two different systems, one including a Support Vector Machine and one with BERT. For the first one, we linguistic pre–processed the dataset to extract hand-crafted features exploited by the classifier. For the second one, we oversampled the dataset to achieve better results. Our results show that the SVM-based system, without the worry of having to oversample, has better performance than the BERT-based one, achieving an F1-score of 91.21%.

1 Introduction

Nowadays, when people want to buy a product or service, they often rely on online reviews of other buyers/users (think of online sales giants like Amazon). Likewise, patients are increasingly relying on reviews on social media, blogs and forums to choose a hospital where to be cured. This behaviour is occurring not only abroad (Greaves et al., 2012; Gao et al., 2012), but also in Italy. This is also demonstrated by the increasing amount of reviews in QSalute[1], one of the most popular Italian ranking websites in healthcare. These reviews are often ignored by hospital companies, which do not exploit the potential of such data to understand patients' experiences and consequently improve their services. Due to the large amount of data, there is a need for automatic analysis techniques. To meet these needs, we decided to introduce a sentiment analysis system based on machine learning techniques, in order to classify whether a review has positive or negative sentiment. Since such systems require annotated data, the first step was to build a brand-new dataset. We present it in the next section. Then, we developed two systems based on two different classifiers described in Section 3 together with the features extracted from the text. In Sections 4 and 5 we show the experiments conducted during this study, the obtained results and their discussion. Finally, the last section provides concluding remarks and some possible future developments. While there exist several works on affective computing in several domains for the Italian language (Basile et al., 2018; Cignarella et al., 2018; Barbieri et al., 2016), at the time we are writing there are no references in literature that address this particular domain in Italian. Thus, for the best of our knowledge, this is the first work of sentiment analysis on Italian reviews in healthcare.

2 Dataset

QSalute is an Italian portal where users share their experiences about hospitals, nursing homes and doctors. We have collected a total of *47,224* documents (i.e. reviews). Each document consists of the free text of the review and other metadata such as the document id, the disease area to which the document belongs and the title. In addition, among the provided metadata there is the average grade, i.e. the mean over the votes in four categories: Competence, Assistance, Cleaning and Services.

In this work, documents with an average grade less than or equal to 2 were assigned to the negative class (-1), while documents with an average grade greater than or equal to 4 were assigned to the positive class (1). The remaining documents were labelled with the neutral class (0). The

Corresponding author: Mario Merone

[1]www.qsalute.it

dataset is strongly unbalanced towards the positive class: 40641 reviews for the positive class, 3898 for the neutral class and 2685 for the negative class. However, in this work, neutral reviews were discarded thus resulting in a dataset composed by *43326* reviews. The following analyses are then referred to this subset: in Table 1 we report some features of the dataset for each site (i.e. the disease area), while the distribution of tokens over their length is reported in Figure 1.

Site	Positive / Total	Lexicon	Overlap(%)
Nervous System	9984 / 10595	34827	69.93
Hearth	5297 / 5491	22677	79.27
Haematology	353 / 377	5336	93.91
Endocrinology	630 / 699	7417	92.40
Endoscopy	1342 / 1484	12046	88.31
Facial	757 / 791	7686	92.13
Genital	2365 / 2552	15605	85.33
Gynaecology	2115 / 2293	14438	90.57
Infections	187 / 220	4001	94.98
Ophthalmology	2167 / 2339	13449	85.43
Oncology	5732 / 6033	25178	79.70
Otorhinology	1156 / 1227	9738	89.91
Skin	763 / 883	8442	90.43
Plastic Surgery	766 / 795	8026	92.04
Pneumology	824 / 982	9454	90.09
Rheumatology	528 / 598	7239	92.14
Senology	3644 / 3783	17497	87.99
Thoracic Surgery	1131 / 1225	10214	90.59
Vascular Surgery	900 / 959	9157	90.05

Table 1: Dataset features for each site. In the first column are reported the name of sites, in the second column are reported the number of positive reviews whit respect to the total numbers of reviews, while in the third one are reported the lexicon values in terms of the number of unique words. Furthermore, in the last column are reported the lexicon overlap (in percentage) of each site with respect to all the others.

The dataset is released on: `www.github.com/lbacco/Italian-Healthcare-Reviews-4-Sentiment-Analysis`.

3 Methods

We developed two systems based on two state-of-the-art classifiers from the state-of-the-art for sentiment analysis, Support Vector Machine and BERT. In this Section, we present the implemented classifiers.

3.1 System 1 based on Support Vector Machine (SVM)

In order to build the first system, we followed the approach proposed by (Mohammad et al., 2013) for the sentiment analysis of English tweets and we adapted it for Italian reviews in healthcare. More precisely, we implemented a Support Vector Machine (SVM) classifier with linear kernel, in terms of liblinear (Fan et al., 2008) rather than libsvm in order to scale better to large numbers of samples, as also reported in the documentation[2] of the LinearSVC model.

Firstly, all documents pass through a preprocessing pipeline, consisting of a sentence splitter, a tokenizer and a Part-Of-Speech (POS) tagger (all of these tools have been previously developed by the *ItaliaNLP*[3] laboratory). Then, documents pass through a step of feature extraction, illustrated in the next section.

3.1.1 Feature Extraction

All features were chosen due to their effectiveness shown in several tasks for sentiment classification for Italian (Cimino and Dell'Orletta, 2016). We refer to these features under the name of *hand-crafted* features and *embedding* features.

Raw and Lexical Text Features

- **(Uncased) Word *n*-grams**: presence or absence of contiguous sequences of *n* tokens in the document text, with *n*={*1, 2, 3*}.
- **Lemma *n*-grams**: presence or absence of contiguous sequences of *n* lemmas occurring in the document text, with *n*={*1, 2, 3*}.
- **Character *n*-grams**: presence or absence of contiguous sequences of *n* characters occurring in the document text, with *n*={*2, 3, 4, 5*}.
- **Number of tokens**: total number of tokens of the document.
- **Number of sentences**: total number of sentences of the document.

Morpho-syntactic Features

- **Coarse-grained Part-Of-Speech *n*-grams**: presence or absence of contiguous sequences of *n* grammatical categories, with *n*={*1, 2, 3*}.
- **Fine-grained Part-Of-Speech *n*-grams**: presence or absence of contiguous sequences of *n* (fine-grained) grammatical categories, with *n*={*1, 2, 3*}.

[2] `www.scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

[3] `www.italianlp.it`

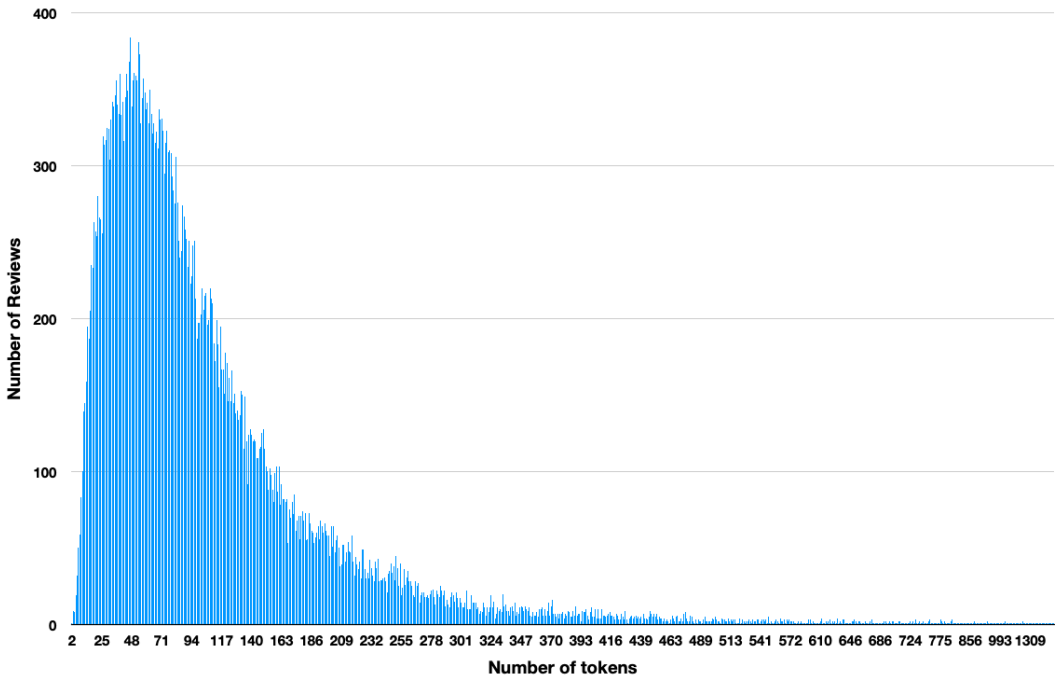

Figure 1: Distribution of documents according to their length, in terms of the number of tokens. The shortest document has only two tokens, while the longest one has 3571 tokens. On average, the reviews are 106.41 tokens long, with a standard deviation of 102.18 tokens.

Word Embeddings Combination: this feature is composed of three vectors. Each vector was calculated by the mean over word embeddings belonging to a specific fine-grained grammatical category: adjectives (excluding possessive adjectives), nouns (excluding abbreviations), and verbs (excluding modal and auxiliary verbs). Word embeddings used in this work are vectors of 128 dimensions, and they were extracted from a corpus of more than 46 million tweets. Such embeddings were already used in (Cimino et al., 2018) and they are available for download at the website of *ItaliaNLP*[4]. Furthermore, three features have been added to indicate the absence of word embeddings belonging to such categories, for a total of 387 $(128 * 3 + 3)$ features.

3.2 System 2 based on BERT

We also implemented Bidirectional Encoder Representations from Transformers, or as better known, BERT, to classify the sentiment of the reviews. BERT is a pre-trained language model developed by (Devlin et al., 2018) at Google AI Language. Pre-trained BERT (available at its GitHub page[5]) may be fine-tuned on a specific NLP task in a specific domain, such as the sentiment analysis for reviews in the healthcare domain. To do that, the original text must be tokenized with its own tokenizer.

4 Experiments

We conducted two types of experiments. In the first one, we wanted to evaluate which of the systems was the best. For each configuration, we have trained and tested the system using a stratified k-fold cross-validation (with $k = 5$). In the second part, we wanted to evaluate the robustness of the best system in a context out-domain, dividing the folders by disease sites. The software has been entirely developed in Python.

4.1 System 1

We tested three different configurations of our SVM-based system, depending on the sets of features used in the experiment: only hand-crafted features (more than 626 thousands features), only embeddings (387 features), and a combination of both. For such experiments, the features that have shown to not bring improvements to the performance (numbers of tokens and sentences), or even

[4] www.italianlp.it/resources/italian-word-embeddings

[5] www.github.com/google-research/bert

to lower it (fg-POS n-grams, Lemmas n-grams with n={*2, 3*}) during a preliminary experimental phase were excluded from the hand-crafted features set. Thus, it turns out that such set is composed only of Uncased Word and cg-POS n-grams with n={*1, 2, 3*}, and Lemmas. In order to reduce the dimensionality of the set, but also to improve the performance of our system, the features pass through a step of filtering after being extracted for the training set. Each feature that appears less than a certain threshold *th* within the training set can be assumed to be not so relevant and is therefore discarded. Such threshold has been set equal to 1 (*th=1*) after a search of the optimal value during the preliminary experimental phase.

4.2 System 2

The experiments with BERT were conducted using the same partition into the *5* folds used during the experiments with the SVM-based classifier. This division allowed us to compare the results achieved by the two classifiers. The BERT model used in our experiments is the multilingual cased pre-trained one.

We tested two different approaches. These experiments have followed two pipelines. In the first one, the model was fine-tuned with folds from the original dataset described in section 2. In the second one, each fold was obtained by oversampling the minority class (i.e. the negative one) in the original fold. The oversampling was obtained by multiplying each negative sample in the fold by 4. These results in the ratio of negative to positive samples being increased from about 1:16 to about 1:4. Other experiments were conducted further increasing the ratio to about 1:2, but this has not led to significant improvements in performance at the expense of computational time. For both the approaches, the model was fine-tuned for 5 epochs on a 12 GB *NVIDIA* GPU with *Cuda 9.0* with the following hyperparameters:

- maximum sequence length of 128 tokens (it seems reasonable since this number is very close to the average length of the documents in the dataset, as reported in Figure1),
- batch size of 24 samples,
- and a learning rate of $5 * 10^{-5}$.

	$F1_{(1)}$ (%)	$F1_{(-1)}$(%)	F1 (%)
SVM			
Hand-crafted	98.90 ± 0.07	82.73 ± 1.03	90.81 ± 0.55
aEmbeddings	96.16 ± 0.15	62.37 ± 0.74	79.27 ± 0.44
Both	**98.94 ± 0.04**	**83.47 ± 0.72**	**91.21 ± 0.48**
BERT			
w/o oversampling	/	/	/
w/ oversampling	98.60 ± 0.04	77.56 ± 0.81	88.08 ± 0.42
Baseline	96.80	0.00	48.40

Table 2: Results of the experiments in the stratified 5-fold cross-validation. Performances are reported in terms of F1-score (%) on each class and the (macro) average between the two. The best results are shown in bold.

5 Results and Discussion

Table 2 resumes the results of the experiments in stratified 5-fold cross-validation. The performances are reported in terms of the macro average of *F1-score*.

After analyzing these results, we took the best model and we used it in the leave-one-site-out cross-validation context to test the reliability of the system in an out-domain (site) problem. These results are resumed in Table 3.

First of all, we can notice that such performances are much higher of the baseline system, i.e. the performance achieved by a hypothetical model that classifies all the samples as belonging to the majority class (that is, the positive class).

Due to the strong dataset imbalance and the low batch size, training BERT without oversampling the dataset leads the system to classify all samples as belonging to the majority class, i.e. the positive class. This leads to often obtain very bad performance, i.e. the baseline performance. Anyway, when this problem does not come up, the classifier shows the lowest value of the F1-score. These results clearly show the difficulties of BERT to deal with unbalanced datasets. Oversampling the minority class has shown to partially cope with such problems, leading to an improvement in terms of repeatability and performance.

For what concerns the experiments with the SVM-based system, they have shown that hand-crafted features have greater relevance for the task than the embedding features. This suggests that the (Italian) healthcare reviews domain may be particularly lexical. Thus, sets of lexical features show better performance than those similarity-based features. However, the resulting best model is the one with both sets of features, outperforming the BERT-based system best configuration by about three percentage points.

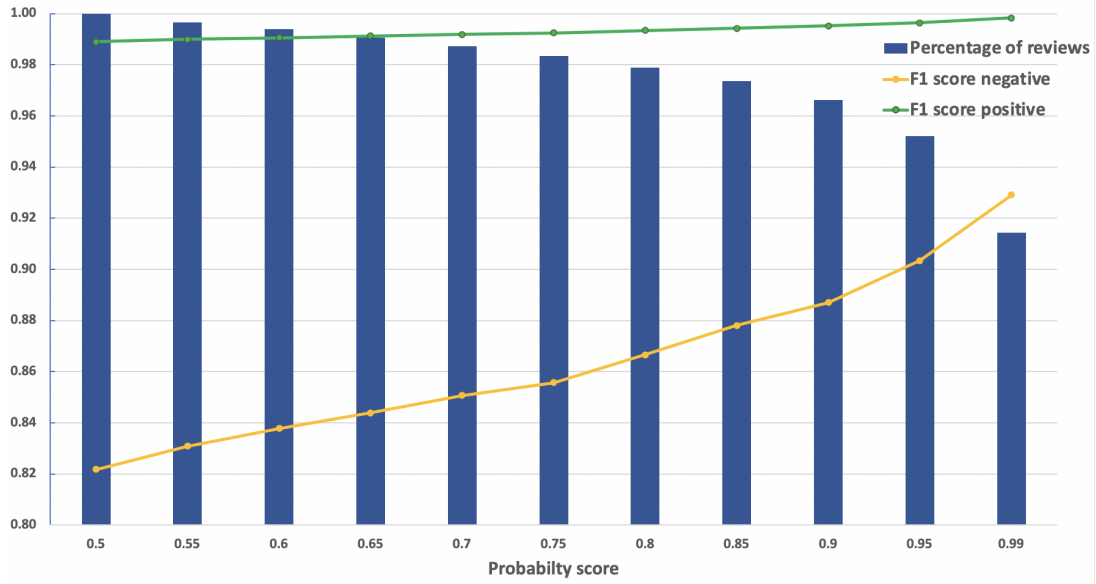

Figure 2: Results in terms of percentage of classified reviews and F1-score over threshold values on the probabilistic score $p \in [0, 1]$ returned by the Platt scaling method applied on top of the SVM-based system. All the results are referred to the k-fold cross-validation (with $k = 5$) fashion. Note that for $threshold = 0.5$, even if the percentage of classified documents is 100%, the value of the macro average of F1-score is lower to the one reported in Table 2. This is due to the inherent inconsistency between the probabilities calculated through the Platt scaling method p and the decision score of the SVM model (i.e. the distance of the sample from the trained boundary, $d \in (-\infty, +\infty)$).

Site	**F1 (%)**	**F1$_{\text{Baseline}}$ (%)**
Nervous System	89.91	48.52
Hearth	90.20	49.10
Haematology	91.10	48.36
Endocrinology	87.79	47.40
Endoscopy	94.34	47.49
Facial	88.31	48.90
Genital	92.12	48.10
Gynaecology	93.64	47.98
Infections	91.09	45.95
Ophthalmology	90.74	48.09
Oncology	90.85	48.72
Otorhinology	89.56	48.51
Skin	93.86	46.35
Plastic Surgery	93.63	49.07
Pneumology	92.76	45.63
Rheumatology	90.75	46.90
Senology	91.29	49.06
Thoracic Surgery	92.62	48.01
Vascular Surgery	90.01	48.41
Average	**91.24**	**47.92**

Table 3: Results of the experiments in leave-one-site-out cross-validation. The first column shows the site used for testing, while the next two columns are the values of performance and baseline in terms of the (macro) average of *F1-score* of each test set.

Furthermore, the leave-one-site-out experiments with this model result in a very good performance, showing the system to be reliable also in an out-domain (site) context. This last result can be due to two factors: 1) the high degree of overlap of the lexicon found in one domain on the lexicon of all other domains; 2) a larger size of the set used for training.

In addition to the two main phases of experiments, we further investigated the confidence of the best model developed in making decisions. The motivation behind this study is that it may have application in real-world cases, where an automated system is required to filter the documents on which it is highly confident (i.e., above a certain threshold) and then passes the most complex documents to a human operator. To do so, we applied the Platt scaling (Platt, 1999) method on top of the trained SVM model. This step is needed to convert the output of the model from a decision score $d \in (-\infty, +\infty)$, i.e. the distance of the test sample from the trained boundary, to a probabilistic score $p \in [0, 1]$, representative of the system confidence in making the decision. Figure 2 resumes the results of this analysis. As expected, the number of documents on which the system makes a decision falls as the confidence threshold required of the system increases. However, this trend does not have such a negative slope and still classify more

than 91% of the documents with 99% confidence. At the same time, the performance advantage is clear, leading to an increase of F1-score on negative samples by more than ten percentage points.

6 Conclusion

In this paper, we have introduced a novel system for sentiment analysis for Italian reviews in Healthcare. For the best of our knowledge, this is the first work of this kind in such domain. To do so, we have collected the first dataset for this domain from the web. Then, we have implemented and compared two types of classifiers of the state of the art for such task, the SVM and BERT. Despite the strong dataset imbalance, we have obtained very good results, especially with the SVM-based system, which outperformed the BERT-based one, while maintaining a low computational burden during training. However, there is a chance that increasing the maximum sequence length of BERT it may outperform our best-developed system. Also, recent work (Nozza et al., 2020) has analyzed the contribution of language-specific models, showing in general improvements over BERT multilingual for a wide variety of NLP tasks. For this reason, it might be worth including in future works the use of specific models for Italian, such as GilBERTo[6], UmBERTo[7], and AlBERTo[8]. The latter was already used for a sentiment classification task (Polignano et al., 2019). Future works on this dataset may also tackle the task of sentiment classification including the neutral class or sentiment regression of the average scores. Moreover, future research may tackle the task of cataloguing reviews to the area of disease they belong, maybe including other features from metadata such as titles.

Acknowledgements

Our work is possible thanks to a general R&D agreement between the National Research Council of Italy (CNR) and Confindustria (the main association representing manufacturing and service companies in Italy) and a specific R&D agreement between Webmonks s.r.l , CNR and the Campus Bio-Medico University of Rome (UCBM).

[6] www.github.com/idb-ita/GilBERTo

[7] www.github.com/musixmatchresearch/umberto

[8] www.github.com/marcopoli/AlBERTo-it

References

[Barbieri et al.2016] Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. 2016. *Overview of the Evalita 2016 SENTIment POLarity Classification Task.*. 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), Napoli, Italy, December 5-7, 2016

[Basile et al.2018] Basile, P., Croce, D., Basile, V., & Polignano. M 2018. *Overview of the EVALITA 2018 Aspect-based Sentiment Analysis Task (ABSITA)*. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.

[Cignarella et al.2018] Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., & Rosso, P. 2018. *Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)*. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.

[Cimino et al.2018] Cimino A., De Mattei L. & Dell'Orletta F. 2018. *Multi-task Learning in Deep Neural Networks at EVALITA 2018*. Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), 86-95.

[Cimino and Dell'Orletta2016] Cimino, A., & Dell'Orletta, F. 2016. *Tandem LSTM-SVM approach for sentiment analysis*. In of the Final Workshop 7 December 2016, Naples (p. 172).

[Devlin et al.2018] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

[Fan et al.2008] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. 2008. *LIBLINEAR: A library for large linear classification*. Journal of machine learning research 9.Aug (2008): 1871-1874.

[Gao et al.2012] Gao, G. G., McCullough, J. S., Agarwal, R., & Jha, A. K. 2012. *A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period*. Journal of medical Internet research, 14(1), e38.

[Greaves et al.2012] Greaves, F., & Millett, C. 2012. *Consistently increasing numbers of online ratings of healthcare in England*. J Med Internet Res, 14(3), e94.

[Mohammad et al.2013] Mohammad S.M., Kiritchenko S., and Zhu X. 2013. *NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets*. In Proceedings of the Seventh international workshop on Semantic Evaluation Exercises, SemEval-2013. 321-327, Atlanta, Georgia, USA

[Nozza et al.2020] Nozza, D., Bianchi, F., & Hovy, D. 2020. *What the [mask]? making sense of language-specific BERT models.* arXiv preprint arXiv:2003.02912.

[Platt1999] Platt, J. 1999. *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in large margin classifiers, 10(3), 61-74.

[Polignano et al.2019] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. 2019. *AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets.* In CLiC-it.