electronics

MDPI

Article

Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian

Rosario Catelli *, Serena Pelosi and Massimo Esposito

Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), 80131 Naples, Italy; serena.pelosi@icar.cnr.it (S.P.); massimo.esposito@icar.cnr.it (M.E.)
* Correspondence: rosario.catelli@icar.cnr.it

Abstract: Recent evolutions in the e-commerce market have led to an increasing importance attributed by consumers to product reviews made by third parties before proceeding to purchase. The industry, in order to improve the offer intercepting the discontent of consumers, has placed increasing attention towards systems able to identify the sentiment expressed by buyers, whether positive or negative. From a technological point of view, the literature in recent years has seen the development of two types of methodologies: those based on lexicons and those based on machine and deep learning techniques. This study proposes a comparison between these technologies in the Italian market, one of the largest in the world, exploiting an ad hoc dataset: scientific evidence generally shows the superiority of language models such as BERT built on deep neural networks, but it opens several considerations on the effectiveness and improvement of these solutions when compared to those based on lexicons in the presence of datasets of reduced size such as the one under study, a common condition for languages other than English or Chinese.

Keywords: sentiment analysis; review; Italian; lexicon; nooj; deep learning; BERT

Citation: Catelli, R.; Pelosi, S.; Esposito, M. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics* **2022**, *11*, 374. https://doi.org/10.3390/electronics11030374

Academic Editor: Maciej Ławryńczuk

Received: 22 December 2021
Accepted: 24 January 2022
Published: 26 January 2022

1. Introduction

The progressive increase in data available to analysts has seen a strong boost in recent years thanks to the growing capillarity with which social networks have spread. One of the major uses of the latter that has contributed to their rise and that of data has been the availability of reviews about commercial products: the possibility of improving products and increasing their visibility or removing them from the market due to a bad reputation have been one of the levers that has moved the interest of market operators towards tools such as sentiment analysis of large amounts of data, in this case review, in order to attract more and more customers [1]. If before the expression of an evaluation of a commercial product was the prerogative of experts in the field through traditional media, now the democratization of this process has brought this possibility to anyone with an internet connection and a way to access it. Consequentially, there has never been such an increase in data and the need for automatic means to extract and analyze the contents in order to modify and direct business strategies in an immediate manner.

In general, the classification of text according to its different aspects has been a focus of research in recent years: opinion mining and sentiment analysis [2,3] through first rule-based systems [4] and then machine [5] and deep [6–9] learning have constituted a continuously improving line of research, thanks to the arrival of increasingly sophisticated language models capable of exploiting prior knowledge and adapting it to the specific tasks for which they have been employed, thus returning better results with less use of computing resources. In detail, language models based on deep neural networks have the advantage of being able to classify the sentiment by learning in an automatic way the key features from the datasets submitted in the training phase, processing sentences with both simple and complex structure; although, however exceptional, these results depend

strongly on the language to be treated and, in particular, on the availability of large datasets on which to train the model beforehand: this situation is common only for English or Chinese languages, while everything else is generally classified as a low-resource language. At this juncture, lexicon-based models are inserted: they exploit pre-constituted dictionaries specific to a language and a domain of interest and are based on formalisms and rules that, although unable to interpret sentences with particularly complex structures, are instead particularly effective in scenarios where the available data are reduced: this argument is particularly valid for a complex but low-resource language such as Italian.

The aim of this paper is to practically compare the performance of these two methods with an ad hoc dataset in Italian language for the task of sentiment analysis, highlighting possible advantages and disadvantages of the two approaches in correspondence of linguistic structures and constructions with specific terminologies, proposing altogether to:

- Verify the performance of one of the best language models available for the Italian language, i.e., $BERT_{Base}$ Italian XXL, by providing a dataset of reviews created ad hoc;
- Test the performance of one of the best NooJ-based lexical analysis systems available for the Italian language, starting with the *Sentix* and *SentIta* lexicons, on the same dataset;
- Understand and compare the performance of the two systems using tools such as SHAP for qualitative analysis and explainability of AI models.

The article is structured as follows: Section 2 describes the background and the most relevant related works, while Section 3 describes tested architectures and experimental setup, used dataset and adopted evaluation metrics. In Section 4 obtained results are discussed and finally in Section 5 conclusions and possible future works are drawn.

2. Background and Related Works

This section reviews the scientific literature concerning the two methodologies compared for the analysis of sentiment within texts: in Section 2.1, the machine- and deep-learning-based methods are discussed, while in the Section 2.2 the lexicon-based methods are illustrated.

2.1. Machine and Deep Learning Based Approaches

In recent years, the continuous expansion of the phenomenon of online reviews has provided a push towards the use of techniques that could automate the process of sentiment analysis, aimed at quickly classifying consumer opinion. The techniques developed have been numerous: starting from the analysis of frequency, role and position of terms in the texts [10] or of specific words and phrases [11], passing through the analysis of syntax [12] and of negations [13], more and more features have been engineered into machine learning algorithms based on the most disparate classifiers, such as maximum entropy and multinomial naïve Bayes [12], but limited by the need for large vocabularies for the training required to extract features for proper classification.

The introduction of the embeddings [14] has been a turning point: through algorithms such as common bag of words or skip-gram it was possible to obtain a vector representation of the tokens constituting the texts providing the context and then predicting the word or vice versa. The limit of this method is related to its static nature: mapped the word in the vector space during the creation of the embedding, the latter remains always the same regardless of the variation of the context of the text in classification and, if the token analyzed was outside the vocabulary of creation then it will not be recognized. Over time, variants of the original algorithm have been proposed, such as the faster GloVe [15] or char2vec [16] based on characters instead of words. Proposals specifically designed for sentiment analysis were also not lacking such as examples of embeddings trained on corpora specifically designed for sentiment analysis [17], then adding the ability to exploit lexical intensity [18] or training on multi-domain scenarios [19]. Readers may find interesting the approaches capable of taking further advantage of coreference and anaphora resolution techniques [20].

Recently, the development of deep neural networks has allowed a further proliferation of embedding techniques [21], as constituents of the first input layer of such networks, such as convolutional, recurrent and transformer-based [22], the foundation of the most modern language models. Convolutional models employ a two-dimensional matrix to represent the generic sentence, such as those proposed by Kim [23] and Kalchbrenner et al. [24]. With the use of specialized convolutional neural networks for the target sentiment, Chen et al. [6] provided further improvement in the area of sentiment analysis. Recurrent models, on the other hand, use memory cells to process the state in which information about incoming and previous tokens is contained. The development of sentiment analysis in relation to such models has typically relied on bottom-up representations of sentences [25], then exploiting long short-term memory (LSTM) networks [26] to mitigate gradient and long-range dependent disappearance issues, thereby achieving important results in correct sentiment recognition [27,28]. At the same time, several language models have begun to emerge, relying on LSTM networks, such as embeddings from language models [29] and universal language model fine/tuning [30]. The great innovation delivered by the language models is the provision of networks provided already pre-trained on huge corpora: this made possible to fine-tune such models using a small amount of task-specific data and much less computational resources. Recently, with the introduction of transformers [31], additional language models have arisen such as generative pre-trained transformer [32] and bidirectional encoder representations from transfomers (BERT) [33], often available in monolingual and multilingual versions between which scientists are trying to understand commonalities and differences [34], also in the field of sentiment analysis related to online reviews [35], but which have shown great performance improvements by overcoming the sequentiality of previous models and introducing operational parallelism through which countless advantages due to context analysis have been shown making the constituent embeddings themselves dynamic.

2.2. Lexicon-Based Approaches

Lexicon-based approaches rely on the assumption that the text semantic orientation is strictly related to the polarity of words and phrases that occur in it. This is related to content words, namely adjectives [36,37], adverbs [38], nouns [39] and verbs [40] and to phrases and sentences that contain them.

Although manually built lexicons are evidently more accurate than the automatically-built ones, especially in cross-domain sentiment analysis tasks, the manual annotation is a costly activity in term of human resources and time [41,42]. This is the cause of the proliferation of studies on automatic polarity lexicons creation and propagation, which perform this task through morphological methods [43,44], by exploiting the semantic relations of thesauri [45–48] and by using co-occurrence algorithms in large corpora [49–51]. Automatically created dictionaries seem to be more unstable, but usually larger than the manually built ones. Size, anyway, does not always mean quality. It is common for these large dictionaries to have scarcely detailed information. Furthermore, a large amount of entries could denote fewer details in description, or, instead, could mean more noise.

Among the most relevant polarity lexicons for the English language SentiWordNet [46] and the SO-CAL dictionary [41] have to be mentioned. SentiWordNet is based on WordNet 2.0 [52] and has been built by automatically associating each WordNet synset to three scores: *Obj* for objective terms, *Pos* and *Neg* for positive and negative terms. Each score ranges from 0.0 to 1.0. The values are determined on the base of the proportion of eight ternary classifiers (with similar accuracy levels but different classification behaviors), that quantitatively analyze the glosses associated with every synset and assign them the proper label. SentiWordNet 3.0 [53] improves SentiWordNet 1.0. The main differences between them are the version of WordNet they annotate (3.0 for SentiWordNet 3.0), and the algorithm used to annotate WordNet, that in the 3.0. version, along with the semi-supervised learning step, also includes a random-walk step that perfects the scores. The SO-CAL dictionary [41], due to the low stability of the automatically generated lexical databases, has been manually

developed by hand tagging, with an evaluation scale that ranged from +5 to −5, the semantically oriented words that have been found into a variety of sources, namely, the multi-domain collection of 400 reviews belonging to different categories [37]; 100 movie reviews from the Polarity Dataset [10,54]; the whole General Inquirer dictionary [55]. The result was a dictionary of 2252 adjectives, 1142 nouns, 903 verbs and 745 adverbs. The adverb list has been automatically generated by matching adverbs ending in *-ly* to their potentially corresponding adjective. Moreover, also a set of multi-word expressions (152 phrasal verbs, e.g., *to fall apart*, and 35 intensifier expressions, e.g., *a little bit*) have been taken into account. In case of overlapping between a simple word (e.g., *fun*, +2) and a multi-word expression (e.g., *to make fun of*, −1) with different polarity, the latter possesses the higher priority in the annotation process.

The largest part of the state of the art works on polarity lexicons for sentiment analysis purposes focuses on the English language. Thus, Italian lexical databases are mostly created by translating and adapting the English ones such as SentiWordNet and WordNet-Affect. Italian polarity lexica that deserve to be mentioned are Sentiment Italian Lexicon [56], also known as Sentix (https://valeriobasile.github.io/twita/sentix.html, accessed on 01 October 2021); SentIta [57]; the lexicon of the FICLIT+CS@UniBO System [58]; the CELI Sentiment Lexicon [59]; the Distributional Polarity Lexicon [60] and SenticNet [61]. Among others, Sentix merged the semantic information belonging to existing lexical resources in order to obtain an annotated lexicon of senses for Italian. Basically, MultiWordNet [62], the Italian counterpart of WordNet [52,63], has been used to transfer polarity information associated to English synsets in SentiWordNet to Italian synsets, thanks to the multilingual ontology BabelNet [64]. The dictionary contains 59,742 entries for 16,043 synsets. SentIta is a semi-automatically built sentiment lexicon that combines polarity and intensity labels that generate an evaluation scale that goes from −3 to +3 and a strength scale that ranges from −1 to +1. In this lexicon, all the adjectives included in the lexical resources of the Italian module of NooJ (https://www.nooj-association.org/resources.html, accessed on 1 October 2021) have been manually annotated with polarity and intensity scores. Afterwards, morphological finite state automata (FSA) have been used to semi-automatically extend the annotation over verbs, nouns and adverbs [65]. The result is a set of dictionaries of more than 20,000 entries, which has recently been enriched by taboo words [66], idioms [67] and emojis [68]. The lexicon of the FICLIT+CS@UniBO System, which has been created for the EVALITA 2014 SENTIPOLC task, includes adjectives and adverbs from the De Mauro-Paravia Italian dictionary and nouns and verbs from the Sentix database. All its lexical items have been classified according to their polarity by the use of the online sentiment analysis API provided by AI Applied (https://ai-applied.nl/text-apis, accessed on 1 October 2021). The CELI Sentiment Lexicon is a sentiment lexicon that contains simple words, multi-words and idioms, annotated with polarity, intensity, emotion and dominance. It is a proprietary resource that CELI sells with a license of use. The Distributional Polarity Lexicon is a large-scale polarity lexicon, which has been automatically created by deriving it through distributional models of lexical semantics, where the polarity of words is derived by sentences annotated with polarity. SenticNet [61] is a knowledge base for concept-level sentiment analysis, freely available also for the Italian language (SenticNet modules are available also for the English, Spanish, Portuguese, Indonesian and Vietnamese languages.), which does not merely use keywords and word co-occurrence counts, but deepens the implicit meaning associated with commonsense concepts by integrating logical reasoning within deep learning architectures. The resource provides semantic annotations associated with 200,000 natural language concepts, including polarity values that go from −1 to +1.

3. Materials and Methods

Hereafter, tested architectures are introduced in Section 3.1 while the exploited dataset and metrics used to evaluate performance are shown in Section 3.2.

3.1. Tested Architectures

Section 3.1.1 describes the deep-learning-based architecture BERT, while Section 3.1.2 illustrates NooJ, a lexicon-based tool.

3.1.1. Deep Learning-Based Architecture: BERT

Widely used in a number of natural language processing tasks such as named entity recognition or sentiment analysis, bidirectional encoder representations from transformers [33] (BERT) forms the basis of several breakthrough language models to date. The ability of bidirectional context analysis, forward and backward, has allowed these kinds of models to specialize their previous knowledge, acquired through a long phase of pre-training on huge corpora, to the specific task under study: if on the one hand, the inner layers of the deep constituent network preserve their generalization capability, on the other hand, the outer layers of the network adapt themselves in a flexible way to the contents examined during the so-called fine-tuning phase that specializes the architecture, producing an ad hoc model.

The knobs through which to intervene in this fine tuning are the so-called hyper-parameters, of which the main ones have been reported in Table 1. In detail there are: the number of hidden layers that make up the encoder transformer also called transformer blocks in numbers equal to 12, then there are the attention heads also called self-attention [31] always equal to 12, then the hidden size of the feed forward networks and the parameter of the maximum length of the input sequence equal to 768 and 512, respectively, and the number of weights that make up the network equal to 110 million (M). Finally, the learning rate, the number of epochs used for fine-tuning and the batch size equal to 0.00001, 5 and 8 in our case, respectively.

Table 1. Hyper-parameters.

Hyper-Parameter	Value
Attention heads	12
Batch size	8
Epochs	5
Hidden size	768
Hidden layers	12
Learning rate	0.00001
Maximum sequence length	512
Parameters	110 M

The tokenization and management of the out-of-vocabulary words is achieved through the WordPieceModel [69,70], which identifies the common sub-words through which the dictionary is built. Instead, the separation between phrases is achieved through the special token [SEP], while the output vector equal to the hidden dimension H with which to represent the entire sequence and provide input to the downstream classifier is represented through the special token [CLS].

The input of the final fully connected classification layer, i.e., the output of the transformers represented by the last hidden layer provided by the first token, is denoted as a vector $C \in \mathbb{R}^H$, while $W \in \mathbb{R}^{KxH}$ is the parameter matrix of the classification layer where K is the number of categories and the probability for each of them is calculated as:

$$P = softmax(CW^T) \quad (1)$$

Instead of employing the loss function *Categorical Cross Entropy*, valid for multi-class classification and provided by default by the BERT model of Hugging Face (https://huggingface.co/transformers/model_doc/bert.html, accessed on 1 October 2021), it was chosen in this case to employ the loss function *Binary Cross Entropy* (BCE) provided by the torch (https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html, accessed on

1 October 2021) library, which is more suitable for the single label prediction case study; however, in order to have more numerical stability, we chose to employ BCE with Logits (https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html, accessed on 1 October 2021) (BCEwL), which combines the BCE with a sigmoid by exploiting the function LogSumExp (https://en.wikipedia.org/wiki/LogSumExp, accessed on 1 October 2021). Given N the batch size, classification using BCEwL can be described as:

$$\begin{aligned} l(x,y) = L = \{l_1,\ldots,l_N\}^T, \\ l_n = -w_n[y_n \cdot \log\sigma(x_n) + (1-y_n)\cdot\log(1-\sigma(x_n))] \end{aligned} \tag{2}$$

Transformer

The most important BERT architectural component is the Transformer [31]. Its operation, starting from **x** and **y** sequences of sub-words, consists in placing the so-called *[CLS]* token before **x**, then after **x** and **y** the so-called *[SEP]*. Hence, the embedding function E and the normalization layer LN contribute to the embedding in this way:

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_{\mathbf{x}}) \tag{3}$$

$$\hat{h}^0{}_{\mathrm{j}+|x|} = E(y_j) + E(j+|x|) + E(1_{\mathbf{y}}) \tag{4}$$

$$\hat{h}_\cdot^0 = Dropout(LN(\hat{h}_\cdot^0)) \tag{5}$$

thus passing through FF, the Feed Forward layer, it happens that M transformer blocks change the embedding, then GELU (the element-wise gaussian error linear units activation function [71]) and MHSA (the multi-heads self-attention function [31]), obtaining:

$$\hat{h}_\cdot^{\mathrm{i}+1} = Skip(FF, Skip(MHSA, h_\cdot^i)) \tag{6}$$

$$Skip(f,h) = LN(h + Dropout(f(h))) \tag{7}$$

$$FF(h) = GELU(h\mathbf{W}_1{}^\top + \mathbf{b}_1)\mathbf{W_2}^\top + \mathbf{b}_2 \tag{8}$$

where $h^i \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{y}|)\times \mathrm{d_h}}$, $\mathbf{W}_1 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_1 \in \mathbb{R}^{4\mathrm{d_h}}$, $\mathbf{W}_2 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_2 \in \mathbb{R}^{4\mathrm{d_h}}$ and the new $\hat{h}_i$ position is:

$$\begin{aligned} [\ldots,\hat{h}_i,\ldots] &= MHSA([h_1,\ldots,h_{|\mathrm{x}|+|\mathrm{y}|}]) \\ &= \mathbf{W}_o Concat(h_i^1,\ldots,h_i^N) + \mathbf{b}_o \end{aligned} \tag{9}$$

In the attention heads, which are N, it happens that:

$$h_i^j = \sum_{k=1}^{|\mathbf{x}|+|\mathbf{y}|} Dropout(\alpha_k^{(i,j)})\mathbf{W}_V^j h_k \tag{10}$$

$$a_k^{(i,j)} = \frac{exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}}{\sum_{k'=1}^{|\mathbf{x}|+|\mathbf{y}|} exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}} \tag{11}$$

where $h_i^j \in \mathbb{R}^{(d_h/N)}$, $\mathbf{W}_o \in \mathbb{R}^{\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_o \in \mathbb{R}^{\mathrm{d_h}}$ and $\mathbf{W}_Q^j$,$\mathbf{W}_K^j$,$\mathbf{W}_V^j \in \mathbb{R}^{\mathrm{d_h/N}\times \mathrm{d_h}}$.

BERT$_{\text{Base}}$ Italian XXL

In order to test BERT on our dataset in the Italian language, the version used was the best one provided through the Hugging Face framework (https://github.com/huggingface/transformers, accessed on 1 October 2021) by the MDZ Digital Library team of the Bavarian State Library (https://huggingface.co/dbmdz/, accessed on 1 October 2021): BERT$_{\text{BASE}}$ Italian XXL. This version of BERT is pre-trained on texts taken from a recent Wikipedia dump plus various text collections from the OPUS (http://opus.nlpl.eu/, accessed on 1 October 2021) corpus plus the Italian OSCAR (https://traces1.inria.fr/oscar/, accessed on 1 October 2021) corpus, for a total of 81 GB of text and 13 billion tokens.

SHAP Explanation Approach

In addition to Accuracy and F_1 score metrics and to better discuss obtained results, the SHAP tool was used. SHAP employs a generic approach to explicate the predictions of any given model: it perturbs model inputs while observing how the output changes [72], based on the idea that the contributions of specific features can be observed by hiding the relevant inputs. In particular, SHAP is based on Shapley's theory of coalition games, in which it calculates the values: these values represent the coalition players, i.e., the features of a data instance, while the prediction represents the payoff of which the fair distribution is established on the basis of these values, which allow us to approximate the number of features in a linear time. Although SHAP was born to explain tabular data and images, it is also well suited for use with language models such as BERT, evaluating the impact of the text fragments that make up an input sentence, i.e., features, on sentiment prediction and their explanation.

3.1.2. Lexicon-Based Method

The lexical method has been tested, in a document-level sentiment classification task, by exploiting a hand-built lexicon, *SentIta*, and an automatically built one, *Sentix*. Regarding the hand-made lexicon, the entries of *SentIta* are labeled with inflectional (FLX) and derivational (DRV) properties and by four sentiment tags: *positive* and *negative* for the property "polarity", and *strong* and *weak* for the "intensity", as can be seen in the example below, with the negative word *sudicio* (English: *filthy*):

sudicio,A+FLX=N106+DRV=SSIMO:N88+POLARITY=NEG+INTENSITY=STRONG

Differently, *Sentix* entries are associated to the part-of-speech (*a* for the adjective), the *WordNet* synset ID, a positive and a negative score from *SentiWordNet*, a polarity score ranging from −1 to 1, and an intensity score ranging from 0 to 1:

sudicio, a, 00419289, 0, 0.75, −1.0, 0.75

In the experiment the words are not considered alone; instead, they are treated into a set of co-occurrence rules that modify their scores according to the syntactic contexts in which they occur. Lexical, morphological and syntactical indicators are the markers that lead the analysis of the polar words in context, for instance:

- Negation, e.g., *niente affatto* (English: *no way*) and *per nulla al mondo* (English: *for anything in the world*);
- Intensification, e.g., *davvero*$^{[+]}$ *eccezionale*$^{[+3]}$ [+3] (English: *truly exceptional*) and *Parzialmente*$^{[-]}$ *deludente*$^{[-2]}$ *anche il reparto degli attori* [−1] (English: *Partially unsatisfying also the actor staff*);
- Comparison, e.g., *Il suo motore era anche il più brioso*$^{[+2]}$ [+3] (English: *Its engine was also the most lively*) and *Un film peggiore di qualsiasi telefilm* [−3] (English: *A movie worse than whatever tv series*).

The contextual shifting of the words has been handled by generalizing all the words endowed with the same prior polarity. Contextual operators, such as negation and comparison markers, intensifiers and down-toners, do not always change a sentence polarity in its positive or negative counterparts, they often have the effect of increasing or decreasing the sentence score, so it is better to talk about valence *shifting* rather than *switching*. The idea is that the final polarity of an expression modified by the context can be modulated by taking into account, at the same time, both the polarity of the opinionated words and the strength of the contextual indicators: Tables 2–4 show examples of co-occurrence rules among negation operators and sentiment words.

Table 2. Negation rules.

Negation Operator	Sentiment Word	Word Polarity	Shifted Polarity
non (not)	fantastico (fantastic)	+3	−1
	bello (beautiful)	+2	−2
	carino (nice)	+1	−2
	scialbo (dull)	−1	+1
	brutto (ugly)	−2	+1
	orribile (horrible)	−3	−1

Table 3. Negation rules with strong operators.

Strong Negation Operator	Sentiment Word	Word Polarity	Shifted Polarity
per niente (in no way)	fantastico (fantastic)	+3	-2
	bello (beautiful)	+2	−3
	carino (nice)	+1	−3
	scialbo (dull)	−1	+2
	brutto (ugly)	−2	+2
	orribile (horrible)	−3	+1

Table 4. Negation rules with weak operators.

Strong Negation Operator	Sentiment Word	Word Polarity	Shifted Polarity
poco (little)	fantastico (fantastic)	+3	−1
	bello (beautiful)	+2	−1
	carino (nice)	+1	−1
	nuovo (new)	0	−1
	scialbo (dull)	−1	+1
	brutto (ugly)	−2	+1
	orribile (horrible)	−3	−1

A network of *local grammars* has been designed on a set of rules that compute the individual polarity scores of words, according to the contexts in which they occur. In general, the sentence annotation is performed using embedded local grammars in the shape of FSA. Local grammars are algorithms that, through grammatical, morphological and lexical instructions, are used to formalize linguistic phenomena and to parse texts. They are defined *local* because, despite any generalization, they can be used only in the description and analysis of limited linguistic phenomena.

NooJ (https://www.nooj-association.org/, accessed on 1 October 2021) is the Natural Language Processing tool used in this work, in the lexicon-based task, for both the language formalization and the corpora pre-processing and processing, at the orthographical, lexical, morphological, syntactic and semantic levels [73]. The NooJ annotations can go through the description of simple word forms, multi-word units and discontinuous expressions. Lexical items and their semantic labels are systematically recalled into local grammars,

which are algorithms in the shape of enhanced recursive transition networks (ERTN) that, through grammatical, morphological and lexical instructions, are exploited in order to formalize linguistic phenomena and to parse texts. The NooJ Finite State Automaton, illustrated in Figure 1, is an abstract device made up by a finite set of states (S) connected by transitions (t), with which it is possible to design a set of patterns able to recognize specific strings. FSA always goes from the initial state (Si) to the final one (Sf). NooJ FSA are special kinds of ERTN that also allow the use of outputs that can describe the recognized patterns, embedded graphs, loops, variables (V) and constraints (C). Furthermore, they can be placed anytime in relation to electronic dictionaries, which can be recalled into each state of the ERTN.

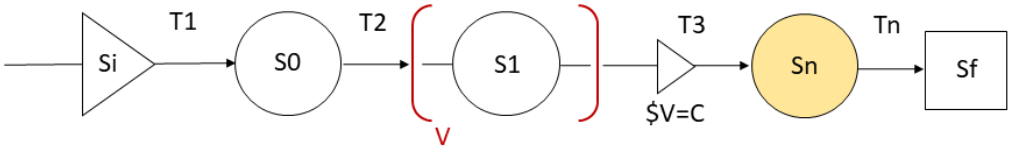

Figure 1. Example of a Nooj Finite State Automaton.

More in detail, with reference to the syntactic treatment of sentiment lexicons, co-occurrence rules have been computed trough the finite-state technology by a network of more than 100 embedded graphs that confer the same polarity values to those expressions in which words belonging to the same classes occur and which are described by the same rules. Such classes of words correspond to the six values that ranges from −3 to +3, that represent the different negative and positive word polarities. There are more than 80 rules and they refer to negation, comparison, intensity and combination of their markers.

All the electronic sentiment dictionaries and the co-occurrence rules used in this paper have been formalized or converted into the NooJ format; therefore, in order to consider both *SentIta* and *Sentix* in context, by recalling them into their syntactic module of NooJ for sentiment analysis, the *Sentix* polarity and intensity scores have been translated into the NooJ labels. Its lemmas have been also enriched with inflectional and derivational properties from the Italian module of NooJ. The purpose was to obtain a version of *Sentix* in the NooJ format that was able to also allow the syntactic treatment of its words. Once the polarities of the dictionaries are extracted and modified, according to their syntactic contexts, the score of the whole review is simply measured by evaluating the arithmetic mean of all the oriented expressions extracted by the tool.

3.2. Dataset and Evaluation Metrics

The dataset exploited in this paper has been built by extracting Italian opinionated documents from e-commerce and opinion websites, such as www.ciao.it accessed on 1 October 2021, www.amazon.it accessed on 1 October 2021, www.mymovies.it accessed on 1 October 2021, www.tripadvisor.it accessed on 1 October 2021. It is composed of 600 reviews (126,184 tokens) about six different products and services, namely cars, smartphones, books, movies, hotels and videogames. Each one of the mentioned category is associated with 50 positive ad 50 negative texts. The distinction among positive and negative reviews is based on the structural data directly selected by the users. This choice poses several challenges which are related to:

- The identification of the proper polarity of the reviews that are close to neutrality;
- The treatment of the reviews that contain both positive and negative claims;
- The correct analysis of reviews in which positive and negative comments are not related to the described product, e.g., delivery issues or plots in movies and books reviews [74].

Furthermore, performances have been evaluated through the following metrics:

- Accuracy, that states the number of labels correctly identified;

- Micro-F_1 score defined as $F_1 \triangleq \frac{2 \cdot P \cdot R}{P+R}$, where Precision $P \triangleq \frac{TP}{TP+FP}$ and Recall $R \triangleq \frac{TP}{TP+FN}$.

4. Results and Discussion

The results of the tested systems are provided in Table 5. The analysis of the numerical results in terms of accuracy and F_1 score easily identifies a winner in Bert, regardless of the lexicon used in combination with the NooJ tool. While it is evident that the language model employed is sufficiently performing even with a modestly sized dataset, on the other hand the difference is not overwhelming and indeed the doubt arises that by improving the lexicons, especially in cases where the data available for fine-tuning are even more scarce, the results of lexicon-based methods may still outperform the more modern deep neural network-based methods. To gain a deeper understanding of performance, for better or worse, of the systems employed, six reviews were analyzed in detail, respectively two not recognized by any system, two recognized only by BERT and two recognized only by NooJ.

Table 5. Accuracy and Micro F_1 results.

	NooJ (SentIta)	NooJ (SentIta + Sentix)	$\text{BERT}_{\text{Base}}$ Italian XXL
Accuracy	0.7583	0.7667	0.9283
F_1	0.8517	0.8846	0.9332

4.1. Quantitative Analysis

The analysis of the errors of the two methods employed in the experiment has been conducted by evaluating the segment-level Precision, Recall, F_1 and Accuracy on the six reviews just mentioned above whose details on performance are reported in Table 6. The analyses made up with both lexicon-based method and BERT have been compared with the evaluation of human annotators over each text portion produced by SHAP, which correspond to the text segments displayed in Tables 7–12. As it can be seen, these segments can be smaller or bigger than a sentence, according to the analyses produced by the tool.

Table 6. Segment-level results.

	NooJ (SentIta)	NooJ (SentIta + Sentix)	BERT
Precision	0.84	0.77	0.73
Recall	0.70	0.85	1
F_1	0.76	0.81	0.84
Accuracy	0.67	0.70	0.73
Wrongly detected reviews			
Precision	0.83	0.79	0.63
Recall	0.67	0.85	1
F_1	0.74	0.81	0.77
Accuracy	0.63	0.71	0.63
Correctly detected reviews			
Precision	0.86	0.67	0.70
Recall	0.75	0.86	1
F_1	0.80	0.75	0.82
Accuracy	0.67	0.60	0.70

The results that describe the analyses of the segments of the six reviews discussed in this section (Table 6) are obviously lower than the ones presented in Table 5, because they regard the error analysis and, for this reason, they focus on the reviews on which the tools, in turn, produced incorrect results. In detail, first rows of Table 6 show the overall performances of the tools on the six review segments. The results related to wrongly and correctly detected reviews, instead, are related to the reviews that have been, respectively,

wrongly or correctly classified as positive or negative in the main task. As an example, wrongly detected reviews for BERT are the ones that have been correctly classified as a whole only by NooJ and the ones on which both the system failed.

Table 7. Neither BERT nor NooJ detect them correctly. Example A.

Review Parts (Italian)	Review Parts (English)	Scores		
		BERT (through SHAP)	NooJ (SentIta)	Nooj (SentIta + Sentix))
"Posizione unica cosa positiva" 2 (su 5 stelle Recensito il 6 settembre 2013 Io e le mie amiche volevamo trascorrere il capodanno a Londra e cercando qualcosa di economico abbiamo trovato il Lonsdale.	"Location only good thing" 2 (out of 5 stars Reviewed 6 September 2013 My friends and I wanted to spend New Year's Eve in London and looking for something cheap we found the Lonsdale.	+0.490	+2 ("positiva", "good") +2 ("economico", "cheap")	+2 ("positiva", "good") +1 ("amiche", friends) +2 ("economico", "cheap")
Dopo aver prenotato abbiamo scoperto che solo le "suite" hanno un bagno privato, mentre per le altre stanze si deve usare il bagno in comune del piano. Il bagno in questione era pulito il primo giorno, ma alla fine del viaggio la vasca era intasata dai capelli degli altri ospiti!	After booking we found out that only the "suites" have a private bathroom, while for the other rooms you have to use the shared bathroom on the floor. The bathroom in question was clean on the first day, but by the end of the trip the tub was clogged with hair from other guests!	+0.323	+2 ("pulito", "clean") −2 ("intasata", "clogged")	−2 ("scoperto", discovered) −1 ("privato", "private") 1 ("privato", "private") −1 ("altre", "other") −2 ("altre", "other") 1 ("comune", "in common") −2 ("comune", "in common") −1 ("piano", "floor") −2 ("piano", "slow") 2 ("pulito", "clean") −1 ("primo", "first") 1 ("primo", "first") −1 ("fine", "end") −2 ("fine", "end") −2 ("intasata", "clogged") −3 ("altri ospiti!", "other guests")
Appena entrati nell'hotel si sentiva una forte puzza, che era presente anche nelle nostre stanze. La mia stanza aveva l'armadio completamente rotto.	As soon as we entered the hotel there was a strong smell, which was also there in our rooms. My room had a completely broken wardrobe.	−0.037	−3 ("completamente rotto", "completely broken")	−1 ("presente", "there") −3 ("completamente rotto", "completely broken") 1 ("completamente rotto", "completely broken") 2 ("completamente rotto", "completely broken")
A causa dell'odore dovevamo dormire con la finestra aperta.	Because of the smell we had to sleep with the window open.	+0.042	0	1 ("aperta", "opened") −2 ("aperta", "opened")
Le lenzuola erano pulite. Ottima la posizione.	The sheets were clean. Good location.	+3.086	+2 ("pulite", "clean") +3 ("ottima")	+2 ("pulite", i.e., "clean") +3 ("ottima", "good location")
Overall score (Predicted label)		>0 (Positive)	>0 (Positive)	<0 (Negative)
Ground Truth		Negative		

Table 8. Neither BERT nor NooJ detect them correctly. Example B.

Review Parts (Italian)	Review Parts (English)	Scores		
		BERT (through SHAP)	NooJ (SentIta)	Nooj (SentIta + Sentix)
Ne sono ancora innamorato anche se la tratto male, (ho delle bruciature di sigarette sui sedili, mi è scappata!).	I'm still in love with it even though I treat it badly (I've got cigarette burns on the seats, it's escaped!).	−0.948	+3 (innamorato, "in love") −2 (male, "badly") −2 (ho delle bruciature di sigarette, "I've got cigarette burns")	−1 ("Ne sono ancora innamorato", "I'm still in love with it") 1 ("Ne sono ancora innamorato", "I'm still in love with it") 1 ("tratto", "treat") 3 ("tratto", "treat2") −2 ("male", "bad") −2 ("ho delle bruciature di sigarette", "(I've got cigarette burns")
difetti: ho il tettuccio apribile: si è rotto dopo i primi caldi e i primi freddi, xkè ci sono dei pezzi di plastica che si seccano dopo un po'.	faults: I have a sunroof: it broke after the first warm and cold spells, because there are pieces of plastic that dry out after a while.	−0.594	−2 ("rotto", "broke")	−2 ("rotto", "broken") −3 ("primi caldi", "first warm") 3 ("primi caldi", "first warm") −3 ("primi freddi", "first cold") 3 ("primi freddi", "first cold")
per fortuna me licambiano adesso in garanzia. Si è rotto anche la serpentina del gasolio, x fortuna in garanzia.	fortunately, they are now replacing it under warranty. The diesel coil also broke, luckily under warranty.	−1.530	−2 ("rotto", "broke")	1 ("adesso", "now") −2 ("rotto", "broken")
Il vetro fischia. il clacson funziona male. lo sterzo è largo, troppo largo. però è bella. motore silenzioso, ma se vai sopra i 100 col tettuccio chiuso, fa rumore. Dietro si tromba bene.	The glass whistles. the horn works badly. the steering is wide, too wide. but it's beautiful. quiet engine, but if you go over 100 with the roof closed, it makes noise. Behind it you fuck well.	−1.520	−2 ("male", "badly") −2 ("troppo largo", "too wide") +2 ("bella", "beautiful") +2 ("bene", "well")	−2 ("male", "bad") −2 troppo largo", "too wide") 2 ("bella", "beautiful") −1 ("silenzioso", "silent") 1 ("silenzioso", "silent") −2 ("silenzioso", "silent") −2 ("chiuso", "closed") 2 ("bene", "well")
Overall score (Predicted label)		<0 (Negative)	<0 (Negative)	<0 (Negative)
Ground Truth		Positive		

What emerged is that, as regards the method based on lexicons, certainly the measurement of the average scores of the words and phrases occurring in the reviews is not suitable for determining their correct orientation. In fact, although there is a very high precision at the segment level for the hand-annotated lexicon *SentIta*, this does not correspond to an adequate performance of the lexicon driven strategy at the document level, when compared to BERT (0.86 *SentIta* and 0.93 BERT). BERT performs significantly better on the segment-level task in the cases in which also the entire documents are classified properly (0.63 of Accuracy in wrongly annotated texts and 0.70 into the correct ones). The main difference between BERT and NooJ is that the fist concentrates the errors on the segments that are labeled by the human annotators as neutral or ambivalent, while the errors of NooJ depend above all on the dictionary features. In fact, *SentIta* presents problems in term of Recall and

Sentix in term of Precision: this relies on the number of entries of the two lexicons and also on the annotation contained in the two resources as detailed in the following paragraph.

Table 9. Only BERT detects them correctly. Example A.

Review Parts (Italian)	Review Parts (English)	Scores		
		BERT (through SHAP)	**NooJ (SentIta)**	**Nooj (SentIta + Sentix)**
Non mi è mai piaciuta.. colpa di quel muso troppo serio. Ha un baule niente male, motori molto tranquilli. E' la classica familiare che strizza l'occhio alle donne…	I've never liked it…it's that too serious nose. It's got a nice trunk, very quiet engines. It's the classic family car that winks at women…	−0.261	−2 ("Non mi è mai piaciuta", "I've never liked it") −2 ("troppo serio", "too serious") +2 ("niente male", "nice") +3 ("molto tranquilli", "very quiet")	−2 ("Non mi è mai piaciuta", "I've never liked it") 3 ("troppo serio", "too serious") −2 ("troppo serio", "too serious") −2 ("niente male", "nice") 2 ("niente male", "nice") 3 ("molto tranquilli", "very quiet") 3 ("classica familiare", "classic family car")
linea tondeggiante, versioni speciali	rounded line, special versions	−0.054	0	0
(come la Pinko, che si distingue per il loro dorato, e la D&G, molto chic che si distingue per la diversa colorazione delle luci posteriori).	(such as Pinko, which is distinguished by its golden colour, and the D&G, very chic which is distinguished by the different colouring of the rear lights).	+0.384	0	3 ("molto chic", "very chic") 1 ("diversa", "different") −2 ("diversa", "different") 2 ("diversa", "different")
Con l'arrivo della nuova C3 la gamma si è ridotta all'essenziale e il modello ha assunto un nuovo nome "C3 Classic". Un usato del 2004/2005 si aggira attorno ai 4800/5000 euro max (esemplari con pochi km e in ottimo stato).	With the arrival of the new C3, the range has been reduced to its essentials and the model given a new name 'C3 Classic'. A used 2004/2005 model is around 4800/5000 euros max (models with few km and in excellent condition).	−0.313	+2 ("essenziale", "essentials") +2 ("esemplari", exemplary") +3 ("ottimo", "excellent")	−2 ("nuova", "new") 2 ("nuova", "new") 1 ("gamma", "range") 2 ("essenziale", "essential") −2 ("nuovo", "new") 2 ("nuovo", "new") 2 ("attorno", "around") 2 ("esemplari", "exemplary") 3 ("ottimo", "excellent")
Purtroppo per i possessori che intendono venderla, l'arrivo del nuovo modello ha fatto scendere parecchio il valore del "vecchio". I consumi? Siamo attorno ai 16 km al litro!	Unfortunately for owners who intend to sell it, the arrival of the new model has caused the value of the 'old' one to drop considerably. Fuel consumption? It's around 16km per litre!	−0.303	0	−3 ("Purtroppo", "Unfortunately") −2 ("nuovo", "new") 2 ("nuovo", "new") −2 ("parecchio", "a lot") −1 ("vecchio", "old") 1 ("vecchio", "old") −3 ("vecchio", "old") −2 ("vecchio", "old") 2 ("vecchio", "old") 2 ("Siamo attorno", "It's around")
Overall score (Predicted label)		<0 (Negative)	>0 (Positive)	>0 (Positive)
Ground Truth		Negative		

Table 10. Only BERT detects them correctly. Example B.

Review Parts (Italian)	Review Parts (English)	Scores		
		BERT (through SHAP)	**NooJ (SentIta)**	**Nooj (SentIta + Sentix)**
La scorsa settimana sono andata da un concessionario Citroen per vedere e provare la nuova C3, praticamente bella. i commenti si sprecano, forse è tato un amore a prima vista, gli interni sono molto belli, e l'esterno ricorda il vecchio maggiolone o comunque macchine di quell'epoca.	Last week I went to a Citroen dealer to see and try out the new C3, which is practically beautiful. The comments are endless, perhaps it was love at first sight, the interior is very nice, and the exterior is reminiscent of the old Beetle or at least cars from that era.	+0.932	+2 ("bella", "beautiful") +3 ("molto belli", "very nice")	−2 ("nuova", "new") 2 ("nuova", "new") −1 ("praticamente bella", "practically beautiful") 1 ("praticamente bella", "practically beautiful") 3 ("molto belli", "very nice") 2 ("esterno", "exterior") −3 vecchio", "old") −2 ("vecchio", "old") −1 ("vecchio", "old") 1 ("vecchio", "old") 2 ("vecchio", "old")
Sicuramente sarà la macchina che acquisterò subito dopo le vacanze estive, visto che ormai la mia piccola utilitaria ha deciso di abbandonarmi.	It will definitely be the car I buy right after the summer holidays, as my little hatchback has now decided to abandon me.	+0.190	0	1 ("Sicuramente", "definitely") 2 ("Sicuramente", "definitely") −2 ("piccola", "small") −1 ("piccola", "small") 2 ("piccola", "small") 2 ("deciso", "decided")
La C3 non è troppo grande, ma nemmeno troppo piccola, l'ideale per una donna,	The C3 is not too big, but not too small either, ideal for a woman,	+0.940	−3 ("troppo grande", "too big") −3 ("troppo piccola", "too small") +2 ("ideale", "ideal")	−1 ("non è troppo grande", " is not too big") −3 troppo piccola", "too small") −2 ("troppo piccola", "too small") 3 ("troppo piccola", "too small") 3 ("ideale", "ideal")
una come me che deve scarrozzare due figli, la spesa da fare, e perchè no, anche per andare a fare shopping con le amiche.	someone like me who has to drive two children, shopping to do, and why not, also to go shopping with friends.	+1.031	0	1 ("amiche") 2 ("amiche")
Consiglio a tutti di andare a vederla e a provarla,	I recommend everyone to go and see it and try it out,	+0.872	0	0
non resterete delusi.	you will not be disappointed.	+0.433	−3 ("delusi", "disappointed")	3 ("delusi", "disappointed")
Overall score (Predicted label)		>0 (Positive)	<0 (Negative)	<0 (Negative)
Ground Truth		Positive		

Furthermore, considering the entire dataset, it must be noticed that the reviews that cause the higher number of errors for both the lexicons in the lexicon-based method (75%) are the positive ones, in the cases in which users discuss both strengths and weaknesses inside them. The presence of negative expressions into positive reviews inevitably make their score decrease and shift towards neutral values. This idea is confirmed by the average absolute values, in the wrongly annotated reviews by both NooJ and Bert, which does not exceed the score of 0.5. Moreover, in the 92% of the NooJ errors and in the 100% of the Bert errors, the average absolute scores are lower than 1.

This is definitely the confirmation that, above all in the case of the lexicon-based method, the scores of the sentences need to be put in relation with one another and with

the semantics of the whole documents. At the same time, it also clear that a classification task can lead to misleading results when the classes can not be sharply divided, because they are poles of a continuum. When the oriented texts are close to neutrality, or they are characterized by ambivalent polarity, the scores attributed by the tool are near to zero and the certainty level of the subjective judges is very low, for both machines and humans.

Table 11. Only NooJ detects them correctly. Example A.

Review Parts (Italian)	Review Parts (English)	Scores		
		BERT (through SHAP)	**NooJ (SentIta)**	**Nooj (SentIta + Sentix)**
Devo ancora capire se mi convince del tutto! Sono molto contenta della maneggevolezza e del motore di questa macchina. E' scattante e ha il cambio preciso preciso…	I still have to see if I'm completely convinced! I am very happy with the handling and the engine of this car. It's quick and has a precise precise gearbox…	+0.933	+2 ("scattante", "quick") +3 ("preciso preciso", "precise precise")	−1 ("ancora", "still") 1 ("ancora", "still") −3 ("molto contenta", "very happy") 3 ("molto contenta", "very happy") 2 ("scattante", "quick") 3 ("preciso preciso", "precise precise")
Però ha alcuni particolari che mi lasciano un po' a desiderare. La visibilità è un po' scarsa sopratutto quando giro a destra la forma rotondetta del montante mi impedisce di vedere bene e devo sempre sporgermi. E' un po' fastidioso!	But it does have a few details that leave me wanting. The visibility is a bit poor, especially when turning right, the rounded shape of the pillar prevents me from seeing well and I always have to lean out. It's a bit annoying!	−1.081	−3 ("po' scarsa", "bit poor") +2 ("bene", "well") −3 ("po' fastidioso", "bit annoying")	−3 ("alcuni particolari", "few details") −3 ("po' scarsa", "a bit poor") 1 ("sopratutto", "above all") 1 ("destra", "right") 2 ("destra", "right") 2 ("bene", "good") −1 ("sempre", "always") −3 ("po' fastidioso!", "a bit annoying")
E poi la grandezza dell'abitacolo per me che sono piccola va benissimo ma quando ci sale qualcuno appena un po' più grosso di me fa fatica a starci comodo con le gambe e le braccia nonostante si sia sistemato il seggiolino.	And then the size of the passenger compartment is just fine for me as a small person, but when someone a little bigger than me gets in, it's hard for me to get my legs and arms into it comfortably, even though I've adjusted the seat.	−0.063	−1 ("fatica", "hard") +2 ("comodo", "comfortably")	2 ("poi") 1 ("benissimo") 3 ("benissimo") 2 ("benissimo") −1 ("fatica") 1 ("fatica") 2 ("comodo")
Nell'insieme sono piccoli particolari. Ma se guidi tutti i giorni contano.	All in all, these are small details. But if you drive every day they count.	+0.105	0	−1 ("insieme") 1 ("insieme") 1 ("particolari") −2 ("particolari")
Overall score (Predicted label)		<0 (Negative)	>0 (Positive)	>0 (Positive)
Ground Truth			Positive	

Table 12. Only NooJ detects them correctly. Example B.

Review parts (Italian)	Review parts (English)	Scores		
		BERT (through SHAP)	**NooJ (SentIta)**	**Nooj (SentIta + Sentix)**
"Ottima posizione" 4 su 5 stelle Recensito il 13 giugno 2013 Albergo veramente carino e moderno nel cuore di piccadilly circus, personale attento e cordiale.	"Great location" 4 out of 5 stars Reviewed 13 June 2013 Really nice and modern hotel in the heart of piccadilly circus, attentive and friendly staff.	+0.265	+3 ("Ottima", "Great") + 2 ("carino", "nice") + 2 ("cordiale", "friendly")	3 ("Ottima", "Great") 2 ("veramente carino", "really nice") −1 ("moderno", "modern") 2 ("moderno", "modern") 2 ("attento", "attentive") 2 ("cordiale", "friendly")
Se soggiornate in questo Hotel vi consiglio di prendere il tè nella libreria, un luogo veramente moderno elegante e rilassante, vi serviranno il tè con una selezione di dolci e sandwich buonissimi!!	If you stay at this Hotel I would recommend you to have tea in the library, a really modern elegant and relaxing place, they will serve you tea with a selection of delicious cakes and sandwiches!!	+0.408	+2 ("elegante", "elegant") +2 ("rilassante", "relaxing") +2 ("dolci", "cakes") +3 ("buonissimi!!", "delicious")	2 ("veramente moderno", "really modern") 2 ("elegante", "elegant") 2 ("rilassante", "relaxing") 2 ("dolci", "cakes") 3 ("buonissimi!!", "delicious")
Devo anche dire che questo Hotel ha due difetti: (1) è leggermente rumoroso (specie se soggiornate durante il weekend si sentono molto i rumori provenienti dalla strada).	I also have to say that this hotel has two flaws: (1) it is a bit noisy (especially if you stay during the weekend you can hear a lot of noise coming from the street).	−0.254	−2 ("rumoroso", "noisy")	1 ("è leggermente rumoroso", "a bit noisy ") 2 ("è leggermente rumoroso", "a bit noisy ") 1 ("specie", "especially")
(2) Durante il nostro soggiorno siamo andati in piscina ma l'acqua non era riscaldata ed era ghiacciata e ci è stato detto che se volevamo avrebbero potuto riscaldare l'acqua per il giorno dopo…	(2) During our stay we went to the swimming pool but the water was not heated and it was freezing cold and we were told that if we wanted they could heat the water for the next day…	−0.652	−2 ("non era riscaldata", "was not heated")	−2 ("non era riscaldata", "was not heated") −2 ("ghiacciata", "freezing") −1 ("detto", "told") 2 ("detto", "told") −2 ("non", "not") 1 ("non", "not")
un albergo di questa categoria non può permettersi uno scivolone del genere!	a hotel of this category can not afford such a slip!	−0.673	0	0
Overall score (Predicted label)		<0 (Negative)	>0 (Positive)	>0 (Positive)
Ground Truth			Positive	

However, these remarks must not be considered drawbacks but research challenges in the sentiment analysis field, which basically motivate the need of a fine-grained textual analysis, that cannot stop at the evaluation of the structured data provided by the users themselves.

4.2. Qualitative Analysis

In order to go in depth in the error analysis, in this paragraph the cases in which the described methodologies fail in the sentiment classification task will be discussed. In this regard, the six emblematic reviews that have been introduced previously have been analyzed in detail. Again, the reference is to those texts which have been improperly labeled by the tools. Both negative and positive incorrect attributions have been taken into

account, referring to the NooJ annotation, to BERT classification, and by both of them. Then, the performances of the tools have been compared with a human reading of such texts.

The first two reviews have been improperly labeled by both the tools (Tables 7 and 8); the third and the fourth have been improperly classified by BERT (Tables 9 and 10) and the latter two refer to NooJ errors (Tables 11 and 12). The main difference between the output of the tools is that Nooj extracts oriented items from the texts, that can be words, multi-word units or phrases, while BERT attributes probabilistic values to entire sentences.

In detail, Tables 7 and 8 refer to reviews that have been wrongly classified by both the methods. Such reviews are made up by positive sentences that open and close the documents and by negative reports in the body of the texts. In the case of Table 7 for a human it is easy to understand that the review is negative, while BERT probably interprets the prominent position of the positive comments as a discourse marker that makes the tool fail in the text classification. The NooJ annotation is negative for SentIta and positive for Sentix, but anyway too close to zero, and consequently to neutrality, to be considered appropriate. This is because of the presence of ambivalent (positive and negative) comments in the same text. Instead, the example reported in Table 8 is difficult to classify also for a human reading, due to the irrational romantic feeling of the author for an almost completely non-functioning car. All the sentences of the text contain dual connotations, which again produce a document-level score close to zero for NooJ and a misleading reading of the sentences for BERT.

Tables 9 and 10 describe the errors made up by NooJ on reviews correctly annotated by BERT. In the first case, again, the presence of both positive and negative expressions brings the average score of Nooj close to zero. In the example reported in Table 10, the main problems with NooJ are related to Recall issues, because the method fails in the extraction of all the oriented expressions actually included in the texts.

Tables 11 and 12, refer to reviews that have been correctly annotated by NooJ, that instead have been cause of errors for BERT. The example of Table 11 is a positive text that is opened and closed by weakly negative comments that probably are considered to be the most relevant ones by BERT and that, at the end, make the system classify the text as negative. The example of Table 12 is basically divided into two parts, the first one is positive and the latter negative, so its classification is challenging for both the methods. NooJ does not fail in the classification because the intensity of the positive items in the first part of the review is higher if compared to the negative ones.

More generally, considering all the reviews included in the error analysis, with regards to the NooJ annotation, the presence of false positive can be noticed, e.g., *dolci* in the last review that is a noun and means *desserts* and not the positive adjective *sweet*, and also of false negatives, e.g., the sentence *un albergo di questa categoria non può permettersi uno scivolone del genere* (English: *a hotel of this category cannot afford one slip like that*).

Overall, it is possible to generalize that the errors caused by BERT seem to be related to the prominence of oriented markers in the text, that can be sometimes misleading. The errors caused by the lexicon methods are mostly related to syntactic issues and to polysemy. As far as syntax is concerned, the case of the long distance between sentiment indicators and contextual modifiers must be mentioned, as happens in *non me la sento di dire che la distribuzione è cosi carente* (English: *I don't feel like saying that the distribution is so lacking*). In this case the negation indicator *non* is ten words away from the polarity word *carente* and the system fails in the evaluation of the negation scope of the adverb. Polysemy pertains, above all, the lexicon that have been built on corpora, *Sentix*, that is by far the one that produces the higher number of errors, despite the fact that it allows the recovery of expressions that were not contained in *SentIta*. In fact, the scores of the entries of *SentIta* are given only if the polarity of the words is evident, also with low intensity. Such resource does not include words and expressions that are too close to neutrality, e.g., *particolare* (English: *characteristic*), words that can change their polarity according to different contexts, e.g., *piccolo* or *grande* (English: *small* or *big*) or domain-dependent polar words, e.g., *silenzioso* (English: *quiet* in the sentence *motore silenzioso*).

Differently, the lemmas from *Sentix*, which are systematically described by their Synonym Set (SynSet), that specify their different usages and shades of meaning, are included in the sentiment lexicon also when their context independent polarity is not so clear. For example, *piccolo* is associated to five different SynSet, four of them with negative scores (*limited in size*, *of little importance* or *influence* or *power*) and one of them with a positive score (*very young*). In contrast, in *SentIta*, words with unpredictable connotations, such as *piccolo*, are not associated with positive or negative polarity scores but treated as downtoners/intensifiers, decreasing/intensifying the intensity of the polarity of the co-occurring words. In the example of Table 10, we see a case of a (negated) negative usage of *piccolo*: *La C3 non è troppo grande, ma nemmeno troppo piccola, l'ideale per una donna* (English: *The C3 is not too big, but not too small either, ideal for a woman*). Neutral (1) and positive (2) usage examples of *piccolo* extracted from the whole text set are reported below:

1. *La colazione e molto buona e suggestiva servita un piccolo salone con caminetto* (English: *The breakfast is very good and evocative served in a small lounge with a fireplace*);
2. *Una piccola bomboniera* (English: *A small party favor*).

This is why *Sentix* has the higher levels of recall and the lower precision, if compared with *SentIta* and BERT. Basically *Sentix* is much larger than *SentIta*, but it contains more noise indeed.

5. Conclusions

This work proposed a comparative study, concerning the ability to analyze sentiment in Italian language on a dataset created ad hoc, between a language model based on deep neural networks such as $\text{BERT}_{\text{Base}}$ Italian XXL and a system based on lexicons such as NooJ. In particular, this study aimed to highlight the limitations and advantages of a language model in contexts other than the English language characterized by the scarcity of datasets such as Italian and for which it may still make sense to rely on models such as those based on the lexicon, deepening with a qualitative analysis based on tools born for the explicability of artificial intelligence such as SHAP.

Based on the results obtained, lexicon-based methods are to be preferred where the datasets are small and the available computational resources limited, under the condition of slightly lower performance. Looking forward, the path of language model-based methods is more attractive: unresolved problems such as the presence of sentiments of different polarities in the same text can be addressed with new implementation solutions by analyzing the limitations of this study that need to be framed in relation to the two proposed approaches.

With respect to the language models, the quality of the results depends both on the model used and on the data; therefore, besides testing the other models available for the Italian language, options could be the development of the dataset used and the utilization of other literature datasets. In addition, the study could be extended to multilingual/crosslingual models and exploit datasets in different languages by testing approaches that leverage transfer learning techniques to face low-resource languages scenarios such as the Italian one. Moreover, it could be useful to expand the classification to a multi-label scenario in which the detection of different emotional states (e.g., anger, happiness, humor, sadness, satire and so on) might weigh differently on the overall sentiment attribution.

On the other hand, with reference to the lexicon-based approach, we basically aim to enrich the syntactic rules for the annotation of the sequences that caused most of the errors discussed in the previous section. In detail, the reported limitation, which can be attributed to syntactic complexity and polysemy, can be addressed by providing grammar networks specialized for the disambiguation of phrases and part-of-speech, based on their textual context. The complexity of FSA will be improved, in order to make them capable of parsing larger phrases and sentences. This way, it is possible to solve a large part of the long-distance dependencies and to attribute a more accurate part-of-speech to the sentiment items, according to their context of occurrence and to their syntactic–semantic relations. Moreover, another drawback, which has been highlighted in the lexical method,

is related to the measurement of the average polarity score of words and phrases located into a review. A possible solution is to include discourse markers in the analyses, in order to evaluate not only each single sentiment marker of a review, but also textual items that express a change of opinion within the review, a synthesis of the orientation expressed by the opinion holder or, for example, a sudden denial of what has been stated previously within the same text.

A very interesting scenario of future work could concern the possible hybridization of symbolic and sub-symbolic methods, so as to have systems able to solicit both the use of lexicons when the resources of pre-training are scarce and the use of previous knowledge when the use case of destination allows it.

Author Contributions: Conceptualization, R.C. and M.E.; data curation, R.C. and S.P.; formal analysis, R.C., S.P. and M.E.; funding acquisition, M.E.; investigation, R.C. and S.P.; methodology, R.C., S.P. and M.E.; project administration, M.E.; resources, M.E.; software, R.C. and S.P.; supervision, M.E.; validation, R.C., S.P. and M.E.; visualization, R.C. and M.E.; writing—original draft preparation, R.C. and S.P.; writing—review and editing, R.C., S.P. and M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Araque, O.; Corcuera-Platas, I.; Sánchez-Rada, J.F.; Iglesias, C.A. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **2017**, *77*, 236–246. [CrossRef]
2. Thet, T.T.; Na, J.; Khoo, C.S.G. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.* **2010**, *36*, 823–848. [CrossRef]
3. Liu, B. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, 2nd ed.; Indurkhya, N., Damerau, F.J., Eds.; Chapman and Hall/CRC: London, UK, 2010; pp. 627–666.
4. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, MI, USA, 1–4 June 2014; Adar, E., Resnick, P., Choudhury, M.D., Hogan, B., Oh, A.H., Eds.; The AAAI Press: Menlo Park, CA, USA, 2014.
5. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2010.
6. Chen, T.; Xu, R.; He, Y.; Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst. Appl.* **2017**, *72*, 221–230. [CrossRef]
7. Pota, M.; Esposito, M.; Palomino, M.A.; Masala, G.L. A Subword-Based Deep Learning Approach for Sentiment Analysis of Political Tweets. In Proceedings of the 32nd International Conference on Advanced Information Networking and Applications Workshops, AINA 2018 workshops, Krakow, Poland, 16–18 May 2018; Barolli, L., Takizawa, M., Enokido, T., Ogiela, M.R., Ogiela, L., Javaid, N., Eds.; IEEE Computer Society: Piscataway, NJ, USA, 2018; pp. 651–656. [CrossRef]
8. Pota, M.; Ventura, M.; Catelli, R.; Esposito, M. An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. *Sensors* **2021**, *21*, 133. [CrossRef]
9. Pota, M.; Ventura, M.; Fujita, H.; Esposito, M. Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Syst. Appl.* **2021**, *181*, 115119. [CrossRef]
10. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, 6–7 July 2002; pp. 79–86. [CrossRef]
11. Mukherjee, S.; Joshi, S. Author-Specific Sentiment Aggregation for Polarity Prediction of Reviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, 26–31 May 2014; Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Reykjavik, Iceland, 2014; pp. 3092–3099.

12. Perikos, I.; Hatzilygeroudis, I. Aspect based sentiment analysis in social media with classifier ensembles. In Proceedings of the 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017, Wuhan, China, 24–26 May 2017; Zhu, G., Yao, S., Cui, X., Xu, S., Eds.; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 273–278. [CrossRef]
13. Diamantini, C.; Mircoli, A.; Potena, D. A Negation Handling Technique for Sentiment Analysis. In Proceedings of the 2016 International Conference on Collaboration Technologies and Systems, CTS 2016, Orlando, FL, USA, 31 October–4 November 2016; Smari, W.W., Natarian, J., Eds.; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 188–195. [CrossRef]
14. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
15. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, 25–29 October 2014; A Meeting of SIGDAT, a Special Interest Group of the ACL; Moschitti, A., Pang, B., Daelemans, W., Eds.; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2014, pp. 1532–1543. [CrossRef]
16. Cao, K.; Rei, M. A Joint Model for Word Embedding and Word Morphology. In Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, 11 August 2016; Blunsom, P., Cho, K., Cohen, S.B., Grefenstette, E., Hermann, K.M., Rimell, L., Weston, J., Yih, S.W., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 18–26. [CrossRef]
17. Li, Y.; Pan, Q.; Yang, T.; Wang, S.; Tang, J.; Cambria, E. Learning Word Representations for Sentiment Analysis. *Cogn. Comput.* **2017**, *9*, 843–851. [CrossRef]
18. Yu, L.; Wang, J.; Lai, K.R.; Zhang, X. Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. *IEEE ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 671–681. [CrossRef]
19. Hao, Y.; Mu, T.; Hong, R.; Wang, M.; Liu, X.; Goulermas, J.Y. Cross-Domain Sentiment Encoding through Stochastic Word Embedding. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 1909–1922. [CrossRef]
20. Sukthanker, R.; Poria, S.; Cambria, E.; Thirunavukarasu, R. Anaphora and coreference resolution: A review. *Inf. Fusion* **2020**, *59*, 139–162. [CrossRef]
21. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
22. Yadav, A.; Vishwakarma, D.K. Sentiment analysis using deep learning architectures: A review. *Artif. Intell. Rev.* **2020**, *53*, 4335–4385. [CrossRef]
23. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, 25–29 October 2014; A Meeting of SIGDAT, a Special Interest Group of the ACL; Moschitti, A., Pang, B., Daelemans, W., Eds.; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1746–1751. [CrossRef]
24. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA, 22–27 June 2014; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2014; pp. 655–665. [CrossRef]
25. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, Grand Hyatt Seattle, Seattle, WA, USA, 18–21 October 2013; A Meeting of SIGDAT, a Special Interest Group of the ACL; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 1631–1642.
26. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Li, D.; Qian, J. Text sentiment analysis based on long short-term memory. In Proceedings of the 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), Wuhan, China, 13–15 October 2016. [CrossRef]
28. Baziotis, C.; Pelekis, N.; Doulkeridis, C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, 16–17 June 2016; Bethard, S., Cer, D.M., Carpuat, M., Jurgens, D., Nakov, P., Zesch, T., Eds.; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2017; pp. 747–754. [CrossRef]
29. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, LA, USA, 1–6 June 2018; Volume 1 (Long Papers); Walker, M.A., Ji, H., Stent, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2227–2237. [CrossRef]
30. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; Volume 1, pp. 328–339. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; The Association for Computational Linguistics: San Diego, CA, USA, 2017; pp. 5998–6008.

32. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. Available online: rhttps://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 October 2021).
33. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
34. Augustyniak, L.; Kajdanowicz, T.; Kazienko, P. Comprehensive analysis of aspect term extraction methods using various text embeddings. *Comput. Speech Lang.* **2021**, *69*, 101217. [CrossRef]
35. Ray, B.; Garain, A.; Sarkar, R. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Appl. Soft Comput.* **2021**, *98*, 106935. [CrossRef]
36. Hatzivassiloglou, V.; McKeown, K.R. Predicting the Semantic Orientation of Adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, 7–12 July 1997; Cohen, P.R., Wahlster, W., Eds.; Morgan Kaufmann Publishers/ACL: Burlington, MA, USA, 1997; pp. 174–181. [CrossRef]
37. Taboada, M.; Gillies, M.A.; McFetridge, P. Sentiment classification techniques for tracking literary reputation. In *LREC Workshop: Towards Computational Models of Literary Analysis*; 2006; pp. 36–43. https://citeseerx.ist.psu.edu/viewdoc/downloaddoi=10.1.1.412.9475&rep=rep1&type=pdf#page=42 (accessed on 1 October 2021).
38. Benamara, F.; Cesarano, C.; Picariello, A.; Recupero, D.R.; Subrahmanian, V.S. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. In Proceedings of the First International Conference on Weblogs and Social Media, ICWSM 2007, Boulder, CO, USA, 26–28 March 2007.
39. Vermeij, M. The Orientation of User Opinions through Adverbs, Verbs and Nouns. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.4909&rep=rep1&type=pdf (accessed on 1 October 2021).
40. Neviarouskaya, A.; Prendinger, H.; Ishizuka, M. Compositionality Principle in Recognition of Fine-Grained Emotions from Text. In Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, CA, USA, 17–20 May 2009; Adar, E., Hurst, M., Finin, T., Glance, N.S., Nicolov, N., Tseng, B.L., Eds.; The AAAI Press: Palo Alto, CA, USA, 2009.
41. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.D.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
42. Bloom, K. Sentiment Analysis Based on Appraisal Theory and Functional Local Grammars. Ph.D. Thesis, Illinois Institute of Technology, Chicago, IL, USA, 2011.
43. Moilanen, K.; Pulman, S.G. The Good, the Bad, and the Unknown: Morphosyllabic Sentiment Tagging of Unseen Words. In Proceedings of the ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, OH, USA, 15–20 June 2008; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2008; pp. 109–112.
44. Neviarouskaya, A. Compositional Approach for Automatic Recognition of Fine-Grained Affect, Judgment, and Appreciation in Text (Soft Computing, < Special Issue> Doctorial Theses on Aritificial Intelligence). *J. Jpn. Soc. Artif. Intell.* **2012**, *27*, 88.
45. Esuli, A.; Sebastiani, F. Determining the semantic orientation of terms through gloss classification. In Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, 31 October–5 November 2005; Herzog, O., Schek, H., Fuhr, N., Chowdhury, A., Teiken, W., Eds.; ACM: New York, NY, USA, 2005; pp. 617–624. [CrossRef]
46. Esuli, A.; Sebastiani, F. Determining Term Subjectivity and Term Orientation for Opinion Mining. In Proceedings of the EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006; McCarthy, D., Wintner, S., Eds.; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2006.
47. Paulo-Santos, A.; Ramos, C.; Marques, N.C. Determining the Polarity of Words through a Common Online Dictionary. In *Progress in Artificial Intelligence, Proceedings of the 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, 10–13 October 2011*; Antunes, L., Pinto, H.S., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 7026, pp. 649–663. [CrossRef]
48. Awadallah, A.H.; Radev, D.R. Identifying Text Polarity Using Random Walks. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; Hajic, J., Carberry, S., Clark, S., Eds.; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2010; pp. 395–403.
49. Qiu, G.; Liu, B.; Bu, J.; Chen, C. Expanding Domain Sentiment Lexicon through Double Propagation. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 11–17 July 2009; pp. 1199–1204.
50. Kanayama, H.; Nasukawa, T. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 355–363.
51. Baroni, M.; Vegnaduzzo, S. Identifying subjective adjectives through web-based mutual information. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), Erlangen, Germany, 7 May 2019.
52. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
53. Baccianella, S.; Esuli, A.; Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17–23 May 2010; Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., Eds.; European Language Resources Luxembourg: Luxembourg, 2010.

54. Pang, B.; Lee, L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004; pp. 271–278. [CrossRef]
55. Stone, P.J.; Hunt, E.B. A computer approach to content analysis: Studies using the General Inquirer system. In Proceedings of the 1963 Spring Joint Computer Conference, AFIPS 1963 (Spring), Detroit, MI, USA, 21–23 May 1963; Johnson, E.C., Ed.; ACM: New York, NY, USA, 1963; pp. 241–256. [CrossRef]
56. Basile, V.; Nissim, M. Sentiment analysis on Italian tweets. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2013, Atlanta, GA, USA, 14 June 2013; Balahur, A., der Goot, E.V., Montoyo, A., Eds.; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2013; pp. 100–107.
57. Pelosi, S. SentIta and Doxa: Italian Databases and Tools for Sentiment Analysis Purposes. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*; Accademia University Press: Turin, Italy, 2015; pp. 226–231. [CrossRef]
58. Di Gennaro, P.; Rossi, A. The FICLIT+CS@UniBO System at the EVALITA 2014 Sentiment Polarity Classification Task. In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014, Pisa, Italy, 9–11 December 2014. [CrossRef]
59. Bolioli, A.; Salamino, F.; Porzionato, V. Social Media Monitoring in Real Life with Blogmeter Platform. In Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), CEUR-WS.org, CEUR Workshop Proceedings, Turin, Italy, 3 December 2013; Volume 1096; pp. 156–163.
60. Castellucci, G.; Croce, D.; Basili, R. A Language Independent Method for Generating Large Scale Polarity Lexicons. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Paris, France, 2016.
61. Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. In Proceedings of the CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, 19–23 October 2020; d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P., Eds.; ACM: New York, NY, USA, 2020; pp. 105–114. [CrossRef]
62. Pianta, E.; Bentivogli, L.; Girardi, C. MultiWordNet: Developing an aligned multilingual database. In *First International Conference on Global WordNet*; Global WordNet Association: Weesp, The Netherlands: 2002; pp. 293–302.
63. Miller, G.A.; Fellbaum, C. WordNet then and now. *Lang. Resour. Eval.* **2007**, *41*, 209–214. [CrossRef]
64. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [CrossRef]
65. Pelosi, S. Morphological Relations for the Automatic Expansion of Italian Sentiment Lexicons. In Proceedings of the Automatic Processing of Natural-Language Electronic Texts with NooJ—9th International Conference, NooJ 2015, Minsk, Belarus, 11–13 June 2015; Revised Selected Papers, Communications in Computer and Information Science; Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 607, pp. 41–51. [CrossRef]
66. Pelosi, S.; Maisto, A.; Vitale, P.; Vietri, S. Mining Offensive Language on Social Media. In Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, 11–13 December 2017; CEUR-WS.org, CEUR Workshop Proceedings; Volume 2006.
67. Pelosi, S. Semantically Oriented Idioms for Sentiment Analysis. A Linguistic Resource for the Italian Language. In Proceedings of the Advanced Information Networking and Applications—Proceedings of the 34th International Conference on Advanced Information Networking and Applications, AINA-2020, Caserta, Italy, 15–17 April 2020; Advances in Intelligent Systems and Computing; Barolli, L., Amato, F., Moscato, F., Enokido, T., Takizawa, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 1151, pp. 1069–1077. [CrossRef]
68. Vitale, P.; Pelosi, S.; Falco, M. #andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline. In Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, 1–3 March 2021; CEUR-WS.org, CEUR Workshop Proceedings; Volume 2769.
69. Schuster, M.; Nakajima, K. Japanese and Korean voice search. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, 25–30 March 2012; pp. 5149–5152. [CrossRef]
70. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
71. Hendrycks, D.; Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. *arXiv* **2016**, arXiv:1606.08415.
72. Kokalj, E.; Skrlj, B.; Lavrac, N.; Pollak, S.; Robnik-Sikonja, M. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, EACL 2021, Online, 19 April 2021; Toivonen, H., Boggia, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 16–21.

73. Bekavac, B.; Kocijan, K.; Silberztein, M.; Sojat, K. (Eds.) Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities. In Proceedings of the 14th International Conference, NooJ 2020, Zagreb, Croatia, 5–7 June 2020; Revised Selected Papers, Communications in Computer and Information Science; Springer: Berlin/Heidelberg, Germany, 2021; Volume 1389. [CrossRef]
74. Maisto, A.; Pelosi, S.; Stingo, M.; Guarasci, R. A hybrid Method for the Extraction and Classification of Product Features from User-Generated Contents. 2017. Available online: http://siba-ese.unisalento.it/index.php/linguelinguaggi/article/download/18352/15749 (accessed on 1 October 2021).