Investigating Embeddings for Sentiment Analysis in Italian*

Giuseppe Gambino[1] and Roberto Pirrone[1 [0000−0001−9453−510X]]

Dipartimento di Ingegneria, Università degli Studi di Palermo, Viale delle Scienze, Edificio 6, 90128, Palermo, Italy
giuseppe.gambino09@community.unipa.it
roberto.pirrone@unipa.it
http://www.unipa.it/dipartimenti/ingegneria

Abstract. The present paper compares the performance of both contextualized and context-free embeddings used for sentiment analysis tasks in Italian. The selected scenario is a pre-analysis stage when the gross architectural parameters of the pipeline have to be devised, while both small data sets can be used for training the model and experiments have to be performed with reduced computational power. Two pipelines have been set up to this aim: the first one makes use of GloVe, which has been suitably trained on the same domain of the task at hand, and a deep neural architecture is used for classification. The second model uses a pre-trained BERT for the Italian language to perform the whole task. The result of our study is that a context-free embedding trained on the task domain outperforms the generic contextualized one. The presented models are reported in detail, along with the experimentations on both the SENTIPOLC 2016 data set and a collection of about 100K TripAdvisor reviews.

Keywords: Sentiment analysis · Text classification · Contextualized word embeddings · Very Deep Convolutional Networks · BERT · GloVe

1 Introduction

The last few years have been a turning point for the development of machine learning models, and in particular of deep learning models. The current positive attitude to share online not only the papers but also their own implementations has brought more and more to a sort of world competition open to all researchers, with the common interest to obtain better results. One of the areas most involved in this phenomenon is the field of Natural Language Processing (NLP) where new studies flourish from day to day. This continuous increase in performances is mainly due to the findings in the field of Distributional Semantics, and in particular to the introduction of contextualized embeddings.

* Supported by PON "Ricerca e Innovazione" 2014-2020, Project "IDHEA" - Innovation for Data Elaboration in Heritage Areas

Despite the high potential of using contextualized embeddings to represent words, these approaches attain their maximum performance if we train them purposely for the task at hand, and very huge data sets have to be used to devise good representations. In this work we present a preliminary study aimed at devising the best word embedding for a NLP task in a pre-analysis scenario when the experiments are constrained by the reduced computational power, while we want to devise an implementation fairly close to the final one because the access to huge computing resources is very limited in general. Also the data sets are reduced versions of the true data to perform a rough tuning of the hyperparameters with the aim of reducing the dimensions of the search space. For the same reasons mentioned above, one prefers to use transfer learning with a pre-trained embedding in place of re-training it from scratch.

To this aim we developed two systems for sentiment polarity classification in Italian to compare the performance of a generic (pre-trained) contextualized embedding against a context-free embedding trained purposely for the task at hand. Due to the limitations posed by the availability of pre-trained embeddings for the Italian, we selected BERT [6], and GloVe [10] respectively for our comparative analysis. Particularly, we used both the data set from the SENTIPOLC 2016 competition [1] and a collection of about 100K Italian reviews from TripAdvisor that we built purposely for this research. Moreover, GloVe embeddings were classified using Very Deep Convolutional Neural Networks [13] in the SENTIPOLC tasks, while Gated Recurrent Units [5] where used in the case of TripAdvisor reviews. We present the implementation details of all the used architectures, and compare the embedding performance. The result of our study is that a suitably trained context-free embedding performs significantly better than a pre-trained contextual one, while consuming low computational resources.

The rest of the paper is arranged as follows. Section 2 reports a brief overview of some of the most recent embedding techniques. In section 3 the structure of the different data sets is addressed along with the detail of the architectures we used. Results are discussed in section 4, and some conclusions are drawn in section 5.

2 Word Embeddings

Both word and character embeddings are a key component of whatever deep learning architecture designed for a NLP task, and they had a tremendous performance increase in the last couple of years since the introduction of the *contextualized embeddings*. The field of distributional semantics starts with the work by Elman [7], while Bengio et al. [4] formulate one of the very first neural language models, which learns embeddings.

Word embeddings enter the NLP arena with Word2Vec [9] where both the Continuous-Bag-Of-Words (CBOW) and Skipgram models are introduced to obtain a dual representation of the words in a text or sentence. Vectors in Word2Vec represent both target words, given a context provided by a suitable window on

both sides of the word itself, and context words given the target one. Although often considered another Word2Vec version, GloVe [10] has a more rigorous approach, which takes advantage of global statistics instead of only local information to create word vectors that capture the meaning in vector space.

Both Word2Vec and GloVe representations provide word representations without considering the true context of the word, apart from the window surrounding the word itself. Contextualized word embeddings fill this gap. Word representations in ELMo [11] are functions of the entire input sentence. It uses a bi-directional Recurrent Neural Network (RNN) using Long Short Term Memory (LSTM) units trained on a specific task to be able to create embeddings. The key behind ELMo is the creation of a language model to be trained for predicting a word in a sequence of words with variable length. The use of a bi-directional LSTM, allows learning a very good representation for each word because all the surrounding text is considered.

The Universal Language Model Fine-tuning for Text Classification [8] or ULMFiT bases its implementation on the concept of transfer learning. It follows that ULMFiT uses a lot of what the model learns during pre-training, more than the other embeddings. This method significantly achieves excellent results on various text classification tasks. Going into details, ULMFit undergoes three phases in its training: a) general-domain language model pre-training where a LSTM based architecture is used to learn the Wikitext-103 corpus, b) target task language model fine-tuning where each layer of the model is tuned with *different* learning rates, and c) target task classifier fine-tuning where *gradual unfreezing* is used that is each layer is unfrozen, and fine tuning proceeds for one epoch, while keeping the others freezed; the process is repeated until convergence.

BERT [6] stands for Bidirectional Encoder Representations from Transformers, and it is the result of combining the concept of bidirectionality introduced by ELMo with the very good results obtained by autoencoders in machine translation thanks to the work by Ashish Vaswani et al. [14]. BERT is categorized as an *autoencoder (AE) language model*. An AE language model aims to reconstruct the original data from corrupted input. BERT works out the directionality constraint of the previous models using a masked language model that randomly masks some of the tokens from the input, and subsequently predicts these tokens based only on its context. BERT performs well in large number of NLP task thanks to its "next sentence prediction" which allows to obtain excellent results in the tasks of natural language inference and paraphrasing, which are based on the prediction of relationships between sentences. BERT offers various pre-trained models available for different languages. We used the Italian version for our experiments.

XLNet [15] is a generalized autoregressive (AR) pre-trained model, which became famous because outperforms BERT in 20 NLP tasks. The idea behind XLNet is not so far from BERT. AR language model is a kind of model using the context words to predict the next word. This type of model is not bidirectional because it can not use both forward and backward context at the same time. To solve this problem in the pre-training phase there is a permutation language

modeling objective that, using permutations on each token of the string, gathers information from all the other ones on both sides. In this way XLNet remedies to the problem of BERT which assumes that masked tokens are independent of each other.

AlBERTo [12]: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets is the latest embedding model available for the Italian language. The creators trained a BERT model for the Italian language; in particular, AlBERTo is pre-trained only on Italian tweets to reach best performances on tasks that concern the Italian language used in social networks. This model obtains the state of the art results in the EVALITA 2016 task SENTIPOLC (SENTIment POLarity Classification). Unfortunaltely, the AlBERTo embedding was not yet available for the experiments at the time of writing the present paper. Actually, it is a very powerful ready to use tool for the Italian language.

3 The Proposed Architectures

As already mentioned above, we built two systems for sentiment polarity classification in both tweets from the SENTIPOLC 2016 competition, and reviews from a collection of TripAdvisor we built purposely. Particularly, we addressed both SENTIPOLC Tasks 1 and 2 that is subjectivity, and polarity classification respectively. Particularly, the goal of Task 1 is identifying the subjectivity of a tweet, with one label that contains 0 for an objective tweet, while a subjective one is labeled with 1. The goal of Task 2 is to identify the polarity of tweet; this is a case of multi-label classification. Indeed, there are two labels; one checks if a tweet is positive or not, and the other one checks if is negative or not. In this manner a tweet can be classified as positive, negative, both positive and negative or without polarity.

We performed just polarity classification for the TripAdvisor reviews, which were split in two classes using the bubble number of each review. In particular we labeles the 5 bubbles reviews as positive, while 1 bubble reviews were set to be negative.

Despite we addressed the same task in both data sets, the language used in these socal media is very different. Tweets are very short, informal language is used very often, while both emoticons and hashtags convey sentiment information. On the other hand, TripAdvisor reviews can be also very long, and are grammatically more correct than tweets. They are very close to the plain text that can be retrieved in the general purpose text corpora as the Wikipedia pages. As a consequence we trained the GloVe embedding differently for the two tasks, and used a different deep neural architecture for classification. In what follows we report in detail the features of each data set, and describe the two architectures.

3.1 Data Sets

SENTIPOLC 2016 Table 1 reports the main features of the SENTIPOLC 2016 competitionn data set, which was given to the participants. As already

pointed out, the length of each tweet is on average 15 words for the training set and 14 for the test set. The other characterizing feature is the presence of many political tweets, that brought us to train GloVe embedding with a dense data set of political tweets.

Table 1. SENTIPOLC 2016 data sets

Data set	Global Tweets #	Political Tweets #	Unique words	Av. length (words)
training	7410	4279	28949	15
test	1998	1498	9771	14

TripAdvisor Dataset The TripAdvisor Dataset[1] is oriented to cultural heritage, and it has been built by scraping the reviews with their label. In this way we have easily obtained a labeled Italian data set. We have only considered sites of cultural interest, and neither restaurants nor hotels. Table 2 reports the main features of such corpus. We scraped 100K reviews, which where used to pre-train the GloVe model. The training/validation set for the classifier is made by 10K reviews equally split in positive and negative ones. The average length of the reviews is 97 words.

Table 2. The TripAdvisor Dataset

Data set	Reviews #	Positive	Unique words	Av. length (words)
TripAdvisor Dataset	10000	5000	23131	97

3.2 Embeddings and Classification

Texts have been pre-processed in the following way. The *ekphrasis* Python library [3] was used to normalize emoticons, url strings, email addresses, Twitter user names, dates and numbers. Moreover, we made all the text lowercase, and removed all the unnecessary white-spaces and symbols. Regarding the hashtags, we have decided not to remove the number sign when training the GloVe embedding to treat each hashtag as a new word thus maintaining all the information around the hashtag, and not around the associated word. A hashtag can convey a sentiment information very different from the word(s) generating it. In general,

[1] TripAdvisor Dataset and the script to generate it is available at the following link: https://github.com/giuseppegambino/Scraping-TripAdvisor-with-Selenium-2019

a hashtag is surrounded by many other ones that could correspond to meaningless sentences if considered as a sequence of plain words. Moreover, a hashtag can be made by many concatenated words (i.e. `#iostoconsaviano`). Again the sentence deriving from the segmentation of composing words does not convey the same sentiment information as the hashtag as a whole.

BERT Embedding The first architecture uses the BERT uncased multilingual version, so it was possible to perform the task on an Italian data sets. The network makes use of a *BERT layer* with three input layers composed as follows:

- input_ids which are just vector representations of words
- segment_ids which are vector representations to help BERT distinguish between paired input sequences
- input_masks to let BERT know that the inputs it is being fed with, have a temporal property masking some of the tokens.

Finally, the last dense layer allows classification thanks to the sigmoid activation function and binary cross-entropy loss function. There are no implementation differences between SENTIPOLC and TripAdvisor data sets applications, except in the size of the input vectors, which have 23 elements in the SENTIPOLC tasks, 300 elements in the TripAdvisor task to accomodate for the longest reviews. Otherwise the implementation is the one recommended by the authors.

GloVe Embedding The GloVe embedding was trained with 230K Italian tweets, whose topic was both generic and political. Since GloVe is unidirectional, a data augmentation technique has been applied, while making different trials with the SENTIPOLC data set. We added all the tweets of the data set in reverse order to the original ones so as to simulate a sort of bidirectionality. Applying this technique we observed improvements in accuracy, while the overfitting was reduced.

The Very Deep Convolutional Neural Network [13] is an implementation that use only small convolutions and pooling operations that works well with a short text like tweets for the nature of convolutional layers. It was used for the SENTIPOLC 2016 tasks. The term "very deep" derives from the high number of convolutional layers, and the authors prove that the performance of this network scales with the depth. We performed fine-tuning of the hyper-parameters for this network. Due to training time constraints, we built a 9-layer VDCNN. The last three layers are dense, and we fixed dropout between them to 0.3. Other parameters are: 32 samples per batch, and the number of filters (64, 128, 256, 512). The implementation for the two task is exactly the same, except for the last classification layer, which used a sigmoidal unit for Task 1 that is a binary classification, while softmax was used for Task 2 for multi-label classification.

The GRU recurrent neural network [5] was used only for the TripAdvisor Dataset. RNNs are best suited for long texts where dependencies between distant words or even between different sentences may convey useful information for the task. Particularly, we found that Gated Recurrent Units performed better than

LSTM cells in our task Fine-tuning of hyper-parameters provided the following values: 32 units, dropout and recurrent dropout fixed to 0.2, 128 samples per batch, 30 training epochs.

4 Results and Discussion

Table 3 and 4 report the performance of our architectures together with the top 5 official results for Task 1, and Task 2 respectively. In particular, each column in both tables reports the F1 score for each label either subjective/objective or positive/negative, and the average F1 score as the overall performance measure to show actually how well the models distinguish the classes. The models are sorted in terms of their average F1 score.

Table 3. The Top 5 F1-scores for SENTIPOLC 2016 Task 1 compared with our architectures

System	Obj	Subj	F1
Unitor.1.u	0.6784	0.8105	0.7444
GloVe-VDCNN	**0,6512**	**0,8248**	**0,7380**
Unitor.2.u	0.6723	0.7979	0.7351
samskara.1.c	0.6555	0.7814	0.7184
BERT	**0,6224**	**0,8100**	**0,7162**
ItaliaNLP.2.c	0.6733	0.7535	0.7134
IRADABE.2.c	0.6671	0.7539	0.7105

Table 4. The Top 5 F1-scores for SENTIPOLC 2016 Task 2 compared with our architectures

System	Pos	Neg	F1
UniPI.2.c	0.6850	0.6426	0.6638
Unitor.1.u	0.6354	0.6885	0.6620
GloVe-VDCNN	**0,6522**	**0,6690**	**0,6606**
Unitor.2.u	0.6312	0.6838	0.6575
BERT	**0,6500**	**0,6523**	**0,6511**
ItaliaNLP.1.c	0.6265	0.6743	0.6504

As already mentioned we tested both GloVe and BERT on our TripAdvisor Dataset. Table 5 shows the F1 score of both architectures. The results show that GloVe performs better than BERT for both Twitter and TripAdvisor datasets. Actually, the GloVe-DCNN architecture ranks second in Task 1, and third in Task 2. We gained several insights from these results.

Table 5. F1-scores of our architectures for sentiment polarity classification in the TripAdvisor Dataset

System	F
Glove-GRU	0,9434
BERT	0,9023

Although BERT is newer than GloVe, and it also includes context analysis, it achieves an almost poor result. We believe that such a behaviour is due mainly to GloVe's Italian-only task-specific pre-training. Even if we collected different tweets from the ones used in the competition, their linguistic features are obviously the same so our GloVe embedding was "focused" in advance to the task. On the other hand, BERT is a multi-lingual model trained on Wikipedia, whose pages are not so sentiment-biased, they use formal grammatical structures, and have much more longer sentences than tweets. The results on the TripAdvisor Dataset support our claim also. In this case BERT does not perform so much worse than GloVe-GRU because the language used in TripAdvisor reviews resembles the one that can be found in Wikipedia.

Going into detail of the results on SENTIPOLC dataset, we noticed that data are biased towards subjective tweets (5000 samples) while the objective ones are 2300. This explains unbalanced results for all the models. It is worth noticing that our architectures rank first, and third respectively on the subjective F1 score. As regards Task 2, GloVe-DCNN beats the winner system for negative polarity tweets. This result is due to the tweets we used for pre-training the embedding. Tweets were collected in a period of political crisis, and are closer to a negative polarity. On the other hand, BERT obtains almost identical values for both labels, and this is due to the use of context that allows the system to exhibit a good discrimination capacity.

Looking at winner systems' implementations, we devised at least two main features in their training that gave them success. The former is distant supervision to increase the size of the training set, and the latter is the use of the TWITA data sets [2], which contain more than 100 millions both generic and topic-specific Italian tweets, and was used to create embeddings. Choosing a not so large data set for training was a precise experimental choice due to the will of achieving complete training using reduced computational resources.

Our objective was obtaining a competitive score with the restriction to have a light and fast implementation. Training was performed on a 2014 MacBook Pro 13" with 8GB RAM, and AVX2 FMA CPU extension. Training GloVe-VDCNN took 4 epochs with 3 minutes per epoch, while BERT implementation was slower than GloVe: the training phase required 3 epochs, and 15 minutes per epoch. A similar behaviour was noted also in the polarity classification task on TripAdvisor reviews. Training GloVe-GRU took 30 epochs, and 1 minute per epoch, while BERT required just 2 epochs, and 50 minutes per epoch.

5 Conclusions

Two neural architectures have been presented in this work, that were aimed at comparing the performance of non purposely pre-trained contextual embeddings with respect to non contextual ones trained with domain-specific data, in a setup with reduced computational sources. The selected task was sentiment classification in Italian social media. The SENTIPOLC 2016 tweets data set, and a purposely built collection of TripAdvisor reviews were used to this aim. We compared a suitably trained GloVe embedding, which was coupled with a deep neural classifier, against a BERT architecture which is available publicly as a multilingual distribution. The result of our investigation is that contextual embeddings perform better than contextual ones, while requiring less computational power to be trained. Results in the SENTIPOLC 2016 competition are satisfactory, and encourages us deepening this issue. Future work will be devoted to try the new AlBERTo embedding model and to use wide data sets to train our embeddings while investigating transfer learning techniques to make our system demanding as less computational resources as possible.

References

1. Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. Overview of the evalita 2016 sentiment polarity classification task. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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.*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.
2. Valerio Basile and Malvina Nissim. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.
3. Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 747–754. The Association for Computer Linguistics, 2017.
4. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.
5. Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.
7. Jeffrey L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225, 1991.
8. Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018.
9. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.
10. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.
11. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018.
12. Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR, 2019.
13. Holger Schwenk, Loïc Barrault, Alexis Conneau, and Yann LeCun. Very deep convolutional networks for text classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1107–1116. Association for Computational Linguistics, 2017.
14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
15. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.