Cimino et al.: Multi-task Learning in Deep Neural Networks at EVALITA 2018 **italian˙twitter˙embeddings**

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. "Multi-task Learning in Deep Neural Networks at EVALITA 2018". In: *EVALITA Evaluation of NLP and Speech Tools for Italian*. Ed. by Tommaso Caselli et al. Torino: Accademia University Press, 2018, pp. 86–95. ISBN: 978-88-319-7842-2 978-88-319-7869-9. DOI: 10.4000/books.aaccademia.4527. URL: https://books.openedition.org/aaccademia/4527 (visited on 06/04/2024).

Annotations: This resource has to be cited in order to use the italian twitter embeddings. In addition to that, it details the system submitted by the ItaliaNLP Lab for the HaSpeeDe task in EVALITA 2018.

Abstract: English. In this paper we describe the system used for the participation to the ABSITA, GxG, HaSpeeDe and IronITA shared tasks of the EVALITA 2018 conference. We developed a classifier that can be configured to use Bidirectional Long Short Term Memories and linear Support Vector Machines as learning algorithms. When using Bi-LSTMs we tested a multitask learning approach which learns the optimized parameters of the network exploiting simultaneously all the annotated dataset labels and a multiclassifier voting approach based on a k-fold technique. In addition, we developed generic and specific word embedding lexicons to further improve classification performances. When evaluated on the official test sets, our system ranked 1st in almost all subtasks for each shared task, showing the effectiveness of our approach.

Commission: The EU Code of conduct on countering illegal hate speech online **european˙commission˙code**

European Commission. *The EU Code of conduct on countering illegal hate speech online*. URL: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (visited on 10/11/2024).

Abstract: The robust response provided by the European Union.

Poletto et al.: Resources and benchmark corpora for hate speech detection **poletto˙resources˙2021**

Fabio Poletto et al. "Resources and benchmark corpora for hate speech detection: a systematic review". In: *Language Resources and Evaluation* 55.2 (June 2021), pp. 477–523. ISSN: 1574-020X, 1574-0218. DOI: 10.1007/s10579-020-09502-8. URL: https://link.springer.com/10.1007/s10579-020-09502-8 (visited on 11/15/2024).

Annotations: Notably, it contains an analysis of lexica developed for hate speech detection.

Abstract: Hate Speech in social media is a complex phenomenon, whose detection has recently gained significant traction in the Natural Language Processing community, as attested by several recent review works. Annotated corpora and benchmarks are key resources, considering the vast number of supervised approaches that have been proposed. Lexica play an important role as well for the development of hate speech detection systems. In this review, we systematically analyze the resources made available by the community at large, including their development methodology, topical focus, language coverage, and other factors. The results of our analysis highlight a heterogeneous, growing landscape, marked by several issues and venues for improvement.

Reynders: 7th evaluation of the Code of Conduct **reynders˙factsheet**

Didier Reynders. *7th evaluation of the Code of Conduct*. Fact-sheet. European Commission, Nov. 2022. URL: https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en?filename=Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf.

Sanguinetti et al.: HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task **haspeede2**

Manuela Sanguinetti et al. "HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task". In: *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Ed. by Valerio Basile et al. Online: CEUR.org, 2020.

Annotations: Contains a description of the task and the datasets. Description and performance of the baselines, overview and evaluation of the submissions to the shared task (this may be interesting both for the methodology and in order to measure ourselves against other submissions).

Sanh et al.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter **distilbert**

Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

Annotations: This resource has to be cited if we use DistilBERT.

Tontodimamma et al.: An Italian lexical resource for incivility detection in online discourses **hurtlex**

Alice Tontodimamma et al. "An Italian lexical resource for incivility detection in online discourses". In: *Quality & Quantity* 57.4 (Aug. 2023), pp. 3019–3037. ISSN: 0033-5177, 1573-7845. DOI: 10.1007/s11135-022-01494-7. URL: https://link.springer.com/10.1007/s11135-022-01494-7 (visited on 11/29/2024).

Annotations: This resource has to be cited in order to use the revised hurtlex lexicon.

Abstract: The exponential growth of social media has brought an increasing propagation of online hostile communication and vitriolic discourses, and social media have become a fertile ground for heated discussions that frequently result in the use of insulting and offensive language. Lexical resources containing specific negative words have been widely employed to detect uncivil communication. This paper describes the development and implementation of an innovative resource, namely the Revised HurtLex Lexicon, in which every headword is annotated with an offensiveness level score. The starting point is HurtLex, a multilingual lexicon of hate words. Concentrating on the Italian entries, we revised the terms in HurtLex and derived an offensive score for each lexical item by applying an Item Response Theory model to the ratings provided by a large number of annotators. This resource can be used as part of a lexicon-based approach to track offensive and hateful content. Our work comprises an evaluation of the Revised HurtLex lexicon.