A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification

Salvatore Carta, Andrea Corriga, Riccardo Mulas, Diego Reforgiato Recupero, Roberto Saia
Department of Mathematics and Computer Science, University of Cagliari, Italy
Artificial Intelligence and Big Data Laboratory (*http://aibd.unica.it/*)
{salvatore,andrea.corriga,diego.reforgiato,roberto.saia}@unica.it,r.mulas1993@gmail.com

Abstract

Nowadays, communications made by using the modern Internet-based opportunities have revolutionized the way people exchange information, allowing real-time discussions among a huge number of users. However, the advantages offered by such powerful instruments of communication are sometimes jeopardized by the dangers related to personal attacks that lead many people to leave a discussion that they were participating. Such a problem is related to the so-called toxic comments, i.e., personal attacks, verbal bullying and, more generally, an aggressive way in which many people participate in a discussion, which brings some participants to abandon it. By exploiting the Apache Spark big data framework and several word embeddings, this paper presents an approach able to operate a multi-class multi-label classification of a discussion within a range of six classes of toxicity. We evaluate such an approach by classifying a dataset of comments taken from the Wikipedia's talk page, according to a Kaggle challenge. The experimental results prove that, through the adoption of different sets of word embeddings, our supervised approach outperforms the state-of-the-art that operate by exploiting the canonical bag-of-word model. In addition, the adoption of a word embeddings defined in a similar scenario (i.e., discussions related to e-learning videos), proves that it is possible to improve the performance with respect to solutions employing state-of-the-art word embeddings.

Motivation and Intuition

The on-line communications between people generate a huge amount of data, giving life to what the researchers defined *Big Data*: a very huge dataset of information from which it is difficult to extract useful information in a reasonable time, if we do not use specific tools and strategies (e.g., *Machine Learning*, *Natural Language Processing*, etc.).

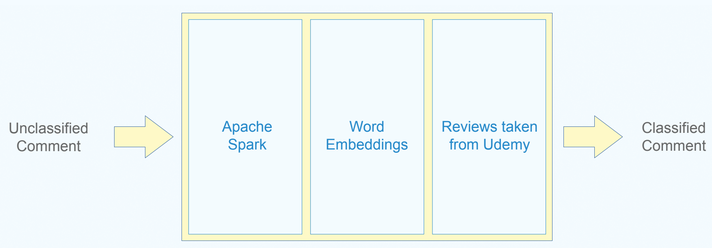
Open Issues

In this scenario the downside is represented by the risks associated with toxic comments, since the advantages related to the aforementioned kind of information (social networks comments, reviews, politic opinions, etc.) are dramatically reduced by those who intervene with verbal attacks, verbal bullying, threats to the person, harassment and, more generally, behaviors that lead some participants to abandon the discussion.

Intuition

Considering that an automatic approach designed to classify the toxicity related to a text must face a multi-class multi-label problem. Our strategy is to use several combinations of word embeddings obtained by exploiting different tools.

Our approach



Metrics

- The first metric we adopted in order to evaluate the performance is the Accuracy, a metric based on the confusion matrix, where *tp*, *tn*, *fp*, and *fn* are, respectively, true positives, true negatives, false positives, and false negatives.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- The second adopted metric is the AUC (Area under the ROC Curve), where given the subsets X and Y , indicates all the possible comparisons between these sub-sets and the result is obtained by averaging all the comparisons.

$$\alpha(X,Y) = \begin{cases} 1, & if\ x > y \\ 0.5, & if\ x = y \\ 0, & if\ x < y \end{cases} \qquad \textbf{AUC} = \frac{1}{|X| \cdot |Y|} \sum_{1}^{|X|} \sum_{1}^{|Y|} \alpha(x,y)$$

Evaluation

Setup and Strategy

We performed two sets of experiments with the aim of comparing several word embeddings approaches, i.e., those based on the state-of-the-art solutions and those built using Udemy data. Moreover, we compared approaches employing word embeddings against the baseline that does not use word embeddings.

Event Classes

▶ **Toxic**	▶ **Obscene**	▶ **Insult**
▶ **Severe Toxic**	▶ **Threat**	▶ **Identity hate**

Experimental Results

Label	**Acc SET$_1$**	**AUC SET$_1$**	**Acc SET$_2$**	**AUC SET$_2$**
toxic	0.81/0.85	0.76/0.73	0.80/0.83	0.74/0.72
severe_toxic	0.85/0.91	0.77/0.76	0.84/0.91	0.73/0.74
obscene	0.84/0.88	0.67/0.74	0.84/0.87	0.66/0.75
threat	0.82/0.90	0.75/0.74	0.81/0.88	0.72/0.75
insult	0.80/0.89	0.74/0.75	0.78/0.88	0.73/0.72
identity_hate	0.84/0.91	0.72/0.75	0.83/0.90	0.70/0.74
Average	**0.83/0.89**	**0.74/0.75**	**0.82/0.88**	**0.71/0.74**

- **Accuracy** and **AUC** are both calculated by using the Term Frequency\Inverse Document Frequency (TF\IDF) function.
- In **SET1**, a logistic regression classifier was trained by using the training set and tested by using the test set.
- In **SET2**, the training and test sets were merged and a k-cross validation with $k = 10$ is adopted, in order to avoid potential biases present in the assigned training and test sets.

Conclusions

- We performed two types of comparisons:
  1. By using the same classification approach with and without the state-of-the-art word embeddings data representation, demonstrating that the word embeddings are able to improve the classification performance with regard to the base-line methods using bag of words models;
  2. By defining word embeddings out of users' comments related to Udemy e-learning platform, with the hypothesis that the creation of word embeddings made by using data closer to the domain taken into account should improve the classification performance.
- The results show how the adoption of word embeddings improve the accuracy with regard to the baselines approaches that do not adopt them. In addition, they prove how the contextual word embeddings outperform the canonical state-of-the-art word embeddings in a specific domain.

Acknowledgements

This research is partially funded by *Italian Ministry of Education, University and Research - Program Smart Cities and Communities and Social Innovation project ILEARNTV (D.D. n.1937 del 05.06.2014, CUP F74G14000200008 F19G14000910008)*. We gratefully acknowledge the support of *NVIDIA Corporation* with the donation of the Titan Xp GPU used for this research.