Extremist Propaganda Tweet Classification with Deep Learning in Realistic Scenarios

Leonardo Nizzoli
University of Pisa and IIT-CNR, Italy
leonardo.nizzoli@iit.cnr.it

Marco Avvenuti
Dept. of Information Engineering, University of Pisa, Italy
marco.avvenuti@unipi.it

Stefano Cresci, Maurizio Tesconi
IIT-CNR, Italy
[name.surname]@iit.cnr.it

ABSTRACT

In this work, we tackled the problem of the automatic classification of the extremist propaganda on Twitter, focusing on the Islamic State of Iraq and al-Sham (ISIS). We built and published several datasets, obtained by mixing 15,684 ISIS propaganda tweets with a variable number of neutral tweets, related to ISIS, and random ones, accounting for imbalances up to 1%. We considered three state-of-the-art, deep learning techniques, representative of the main current approaches to text classification, and two strong linear machine learning baselines. We compared their performance when varying the composition of the training and test sets, in order to explore different training strategies, and to evaluate the results when approaching realistic conditions. We demonstrated that a Recurrent-Convolutional Neural Network, based on pre-trained word embeddings, can reach an excellent F1 score of 0.9 on the most challenging test condition (1%-imbalance).

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Neural networks**.

KEYWORDS

Extremist propaganda; artificial neural networks; cyber intelligence; Twitter

ACM Reference Format:
Leonardo Nizzoli, Marco Avvenuti, and Stefano Cresci, Maurizio Tesconi. 2019. Extremist Propaganda Tweet Classification with Deep Learning in Realistic Scenarios. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292522.3326050

1 INTRODUCTION

The volume and variety of the Twitter stream ensure a comfortable hideout to malicious propaganda activities. Hence, the Social Media Intelligence (SOCMINT) research community has devoted increasing efforts in developing automatic methods to filter malicious tweets [10]. If literature provides several contributions on this topic [6], the proposed techniques are generally tested and evaluated in laboratory conditions [1, 2, 7]. Therefore, stakeholders cannot estimate how those models would perform in the wild.

Motivation. In this work, we focused on the Islamic State of Iraq and al-Sham (ISIS). We addressed the needs of a potential stakeholder who leverages the Twitter Streaming API [4] to obtain a mixture of ISIS propaganda tweets (pro-ISIS), tweets reporting ISIS related content but not supporting the organization (about-ISIS), and completely random ones, with unknown relative proportion. The stakeholder is interested in choosing the suitable state-of-the-art text classification technique, together with the optimal training strategy, to separate pro-ISIS tweets from the others, and in having an estimation of the expected performance.

Contribution. We selected three state-of-the-art, deep learning techniques, representative of the main current approaches to text classification, covering both character- and word-level text representations, and convolutional and recurrent architectures. For comparison purposes, we also included two strong linear baselines. Starting from 15,684 pro-ISIS tweets, we built several datasets accounting for a wide range of positive class imbalances (up to 1%). The results of this extensive investigation addressed the objectives of identifying a good candidate technique, together with a proper training strategy, and of estimating the performance in realistic conditions. In particular, we obtained an excellent F1 score of 0.9 under the most challenging condition of 1%-imbalanced test set. As a key contribution, we published the experimental datasets.

2 MATERIALS AND METHODS

Datasets. In order to reproduce the features of a realistic scenario, we built several datasets by combining pro-ISIS, about-ISIS and random tweets, with different proportions. The 15,684 pro-ISIS English tweets were sampled from a large, reliable, publicly available dataset[1], published by *Fifth Tribe*, a digital agency providing services to US government. We sampled the about-ISIS tweets from a dataset[2] including tweets containing at least one ISIS related keyword. We collected random tweets via the Twitter Streaming API, by means of the *GET statuses/sample* method, which returns a small random sample of all public statuses. The largest and most imbalanced (1%) dataset included, for each pro-ISIS tweet, 6 about-ISIS and 93 random tweets. Less imbalanced datasets (from 2% to 10%, and balanced) were derived from more imbalanced ones, by removing negative instances. Each dataset underwent stratified splitting to create training (64%), validation (16%) and test sets (20%). We published the datasets[3] for reproducibility and further research.

[1] https://www.kaggle.com/fifthtribe/how-isis-uses-twitter
[2] https://www.kaggle.com/activegalaxy/isis-related-tweets
[3] http://ci.iit.cnr.it/ept. For the login credentials, please e-mail leonardo.nizzoli@iit.cnr.it.

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06.
https://doi.org/10.1145/3292522.3326050

Learning techniques. Deep learning, based on Artificial Neural Networks (ANNs), represents the current state-of-the-art for text classification. The main advantages are higher performance, automatic feature extraction and higher generalizability [5]. Limiting ourselves to the state-of-the-art, we identified a possible taxonomy as follows: (i) Convolutional (CNN) vs. Recurrent-Convolutional (RCNN) Neural Network architectures and (ii) character- vs. word-based text representations. CNNs result very effective in capturing the text semantic, whereas RCNNs, due to their sequential architecture, excel in capturing contextual information and long-term semantic features. Characters-based representations enable to learn multi-language models and abnormal character combinations (very common in tweets [3]). Instead, word-based representations commonly leverage word embeddings to incorporate distributional information about words, learned in large text corpora. Due to the different pros and cons of each approach, we decided to compare: (i) a character-based CNN model (`ChCNN`), proposed in [12]; (ii) a RCNN (`WeRCNN`), combined with max pooling [8] and based on pre-trained FastText word embeddings [9]; (iii) an evolution of the previous one (`We2DRCNN`), in which convolution and max pooling were carried out in a two-dimensional arrangement [13], and (iv) two linear baselines (SVM), trained on bag-of-character (`ChSVM`) and bag-of-word (`WoSVM`) n-grams [11].

.

3 RESULTS AND DISCUSSION

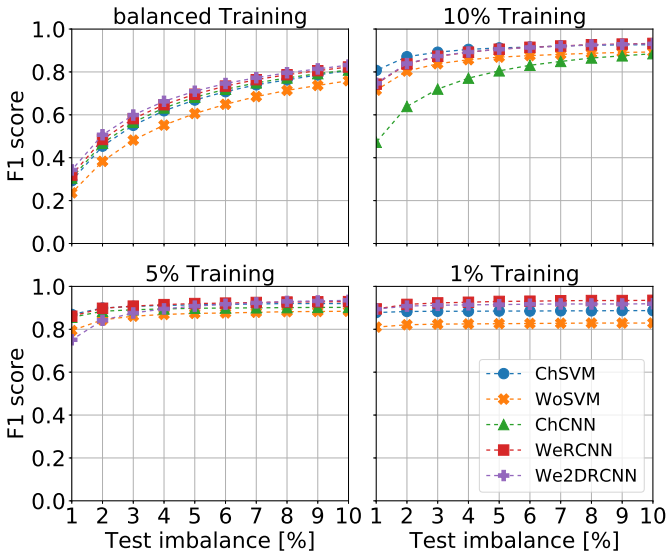

Figure 1: F1 score trends, for different training strategies, when varying the test set imbalance. On 1% training, `ChCNN` performed like a majority classifier (undefined F1 score).

We trained models on balanced, 10%, 5%, 1% training sets, and we evaluated their performance when varying the imbalance of the test sets in the range [1%, 10%]. Figure 1 shows the obtained F1 score trends. Models trained on the balanced training set suffered a serious performance worsening when decreasing the percentage of positive instances in the test set. Instead, the performance showed a more stable trend when we increased the number of negative examples in the training set. The linear baseline `ChSVM` obtained the highest F1 score when using the 10% training set, but it was outperformed on the 1%. `ChCNN` was competitive with the 5% training set, but resulted in a majority classifier, with an undefined F1 score, on the 1% one. Word-based RCNNs clearly outperformed the other candidates on the most imbalanced 1% dataset, keeping very stable on the whole investigated range. In particular, `WeRCNN` obtained the highest F1 score (0.895) on 1% test set.

Fig. 1 trends address the needs of the potential stakeholders, as summarized in Section 1. In fact, they outline RCNNs, based on pre-trained word embeddings, as suitable techniques. Moreover, they support the choice of highly imbalanced training set if an unknown, high imbalance is expected in the real tweet stream. Finally, they allow a performance estimation over a wide range of conditions.

4 CONCLUSIONS

We outlined a framework for solving the task of extremist propaganda tweet classification in realistic scenarios. We provided the performance trends of several text classification techniques, when varying the training strategies and the test conditions. Those trends address the needs of possible stakeholders seeking for a suitable text classification technique, an efficient training strategy and an estimation of the performance expected in the wild. Focusing on the most challenging test condition (1% imbalance), we demonstrated that Recurrent Convolutional Neural Network, based on pre-trained word embeddings, reached a F1 score as high as 0.9 when trained with the same imbalance. Moreover, the measured trends provide a performance estimation on a wide range of conditions. Finally, the dataset were published for reproducibility and further research.

REFERENCES

[1] Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. In *ICDCIT'15*. Springer, 431–442.
[2] Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, and Nico Prucha. 2015. Detecting jihadist messages on twitter. In *EISIC'15*. IEEE, 161–164.
[3] Marco Avvenuti, Stefano Cresci, Leonardo Nizzoli, and Maurizio Tesconi. 2018. GSP (Geo-Semantic-Parsing): Geoparsing and Geotagging with machine learning on top of linked data. In *ESWC'18*. Springer, 17–32.
[4] Stefano Cresci, Salvatore Minutoli, Leonardo Nizzoli, Serena Tardelli, and Maurizio Tesconi. 2019. Enriching Digital Libraries with Crowdsensed Data. In *IRCDL'19*. Springer, 144–158.
[5] Tiziano Fagni, Leonardo Nizzoli, Marinella Petrocchi, and Maurizio Tesconi. 2019. Six Things I Hate About You (in Italian) and Six Classification Strategies to More and More Effectively Find Them. In *ITASEC'19*.
[6] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In *SOCINFO'16*. Springer, 22–39.
[7] Andrew H Johnston and Gary M Weiss. 2017. Identifying Sunni extremist propaganda with deep learning. In *SSCI'2017*. IEEE, 1–6.
[8] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI'15*, Vol. 333. 2267–2273.
[9] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *LREC'18*. ELRA.
[10] David Omand, Jamie Bartlett, and Carl Miller. 2012. Introducing social media intelligence (SOCMINT). *Intelligence and National Security* 27, 6 (2012), 801–823.
[11] Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL'12*. 90–94.
[12] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS'15*. 649–657.
[13] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *COLING'16*. 3485–3495.