No Place For Hate Speech @ HaSpeeDe 2: Ensemble to Identify Hate Speech in Italian

Adriano dos S.R. da Silva
Schoool of Arts, Sciences and Humanities – University of Sao Paulo
Sao Paulo - Brazil
`adriano.santos.silva@usp.br`

Norton T. Roman
Schoool of Arts, Sciences and Humanities
University of Sao Paulo
Sao Paulo - Brazil
`norton@usp.br`

Abstract

English. In this article, we present the results of applying a Stacking Ensemble method to the problem of hate speech classification proposed in the main task of HaSpeeDe 2 at EVALITA 2020. The model was then compared to a Logistic Regression classifier, along with two other benchmarks defined by the competition's organising committee (an SVM with a linear kernel and a majority class classifier). Results showed our Ensemble to outperform the benchmarks to various degrees, both when testing in the same domain as training and in a different domain.

Italiano. *In questo articolo, ci presentiamo i risultati dell'applicazione di un modello di Stacking Ensemble al problema della classificazione dei discorsi di incitamento all'odio nel compito A di EVALITA (HaSpeeDe 2). Il modello è stato quindi confrontato con un modello di regressione logistica, insieme ad altri due benchmark definiti dal comitato organizzatore della competizione (un SVM con un kernel lineare e un classificatore di classe maggioritaria). I risultati hanno mostrato che il nostro Ensemble supera i benchmark a vari livelli, sia durante i test nello stesso dominio di sviluppo che in un dominio diverso.*

1 Introduction

Social networks are already part of people's lives, generating thousands of publications on a daily basis. Even though most of this material presents no real harm to other people, some of it bears discriminating discourse, not rarely filled with hate for minorities or people with different viewpoints.

Defined as "language which attacks or demeans a group based on race, ethnic origin, religion, gender, age, disability, or sexual orientation/gender identity" (Nobata et al., 2016), hate speech represents a problem that cannot be allowed to grow, under the risk of having it lead to more concrete actions, by some people, with truly undesired results.

This is so much of an issue, that some companies have already decided to stop advertising on Facebook[1], for example, as a way to try to pressure the company into facing this problem. Some initiatives have also emerged in order to monitor and combat this type of content, such as the code of conduct that has been signed by some companies (YouTube, Facebook, Twitter) so that this type of publication can be monitored and removed within 24 hours[2].

Due to the large volume of data, machine learning techniques, along with natural language processing, are being used to automate this activity and identify this type of speech more accurately. Other initiatives include the setting up of competitions, aimed at developing and testing different ways to tackle the problem.

One such competitions is the evaluation campaign of Natural Language Processing and Speech Tools for Italian (EVALITA), which started in 2007 aiming at promoting the development and dissemination of language resources for Italian. In its 2018 edition, a task (HaSpeeDe) was proposed to identify hate speech on Facebook and Twitter (Bosco et al., 2018). HaSpeeDe had the par-

[1] `https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html`

[2] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

ticipation of several teams and promising results were presented that stimulated the development of the second edition of the event (HaSpeeDe2) at EVALITA 2020 (Sanguinetti et al., 2020; Basile et al., 2020). In this work, we describe our attempt to deal with the hate speech identification problem HaSpeeDe 2, by developing a stack ensemble of three machine learning models to this task. Weak classifiers used in the ensemble were an SVM with RBF kernel, a Bernoulli Naïve Bayes (NB), and a Random Forest model (RF), with a Linear Regression (LR) model serving as meta-classifier.

For the sake of comparison, and as a way to define some benchmarks to our model, we also developed and tested a Linear Regression classifier, with L2 regularisation, along with both models suggested by HaSpeeDe 2 organising committee, to wit, an SVM model with a linear kernel and a majority class classifier. As it will be made clearer in the forthcoming sections, with a Macro F1-score of 0.749, our ensemble outperforms all benchmarks, for both in and out-of-domain test sets, even though sometimes differences were not high.

The rest of this article is organized as follows. Section 2 presents some related work, aiming at identifying hate speech. Section 3, in turn, gives an overview of HaSpeeDe 2 task. Next, in sections 4 and 5 we explain the preprocessing we made, along with the classifiers we built for this task. Section 6, in turn, presents our results, which are further discussed in Section 7. Finally, Section 8 presents our final considerations to this work.

2 Related Work

Several strategies have been used to identify hate speech. Some classic algorithms, like Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR) and ensemble with these techniques have also shown good results (*e.g.* (Basile et al., 2019; Saha et al., 2018; Malmasi and Zampieri, 2018)).

An SVM with RBF kernel, for example, was used to identify hate speech against immigrants and women in tweets written in English. Achieving a macro-averaged $F1$ score of 0.65 this model was the winner at SemEval 2019 (Basile et al., 2019).

Logistic Regression was another classic model to be applied to hate speech identification in English, in this case focusing in hate speech towards women, with a reported accuracy of 0.70 (Saha et al., 2018). Delivering an accuracy value of 79.8, an ensemble associated with a meta-classifier was also found to perform well in the task (Malmasi and Zampieri, 2018).

With an overall performance of $F1 = 0.749$, our ensemble method looks competitive, when compared to these models. Even though one cannot really make a true comparison between them, we believe this to be an alternative to be considered.

3 Task

HaSpeeDe 2 Task A consists of a binary classification to identify the presence or absence of hate speech in tweets written in Italian. The competition's organising committee provides participants with a data set for training and testing competing models. This data set is slightly imbalanced, with approximately 40% of tweets presenting hate speech language, as shown in Table 1.

Table 1: Data set class distribution

Hate Speech	Not Hate Speech	Total
2766	4073	6839

This data set is supposed to be used by the competition participants to train and test their models. Competing models will then be evaluated in a separate data set, which consists of in-domain and out-of-domain data, defined by the competition's organisation.

4 Preprocessing

As a preprocessing step, we removed stopwords using the NLTK (Natural Language Toolkit [3]) library. For each tweet in the corpus, we also added the following new features:

- The number of words in the tweet;
- The number of exclamation points ('!') present in the tweet; and
- The presence or not of a question mark ('?') in the tweet.

As a final measure, all features related to the tweet's text were normalised in the range between 0 and 1.

[3] `https://www.nltk.org/`

Table 2: Results of the classifiers in the training stage in terms of F1

		Without Preprocessing		**With Preprocessing**	
Classifier	**Lang. Model**	**No Norm.**	**TF-IDF**	**No Norm.**	**TF-IDF**
RF	3-Gram	0.662	0.657	06687	0.667
RF	4-Gram	0.683	0.694	0.690	0.689
RF	5-Gram	**0.701**	**0.701**	0.687	0.686
LR	3-Gram	0.681	0.703	0.676	0.696
LR	4-Gram	**0.711**	0.701	0.706	0.697
LR	5-Gram	**0.711**	0.673	0.708	0.673
NB	3-Gram	0.679	0.679	0.681	0.681
NB	4-Gram	0.689	0.689	**0.694**	**0.694**
NB	5-Gram	0.654	0.654	0.668	0.668

Table 3: Results of the classifiers in the test stage in terms of F1

		Without Preprocessing		**With Preprocessing**	
Classifier	**Lang. Model**	**No Norm.**	**TF-IDF**	**No Norm.**	**TF-IDF**
RF	3-Gram	0.650	0.668	0.650	0.674
RF	4-Gram	0.693	0.694	**0.710**	0.696
RF	5-Gram	0.707	0.709	0.703	0.700
LR	3-Gram	0.675	0.701	0.675	0.709
LR	4-Gram	0.684	0.696	0.685	**0.710**
LR	5-Gram	0.669	0.665	0.707	0.680
NB	3-Gram	0.696	0.696	0.707	0.707
NB	4-Gram	0.718	0.718	**0.740**	**0.740**
NB	5-Gram	0.658	0.658	0.687	0.687

5 Classifiers

In the sequence, three individual classifiers were developed using the Python Sklearn[4] library. These were a Naïve Bayes (NB) with Bernoulli distribution, Logistic Regression (LR) with L2 regularization, and Random Forest (RF) with 150 trees. Each classifier was tested with N-Gram representations (N ranging from 3 to 5), with and without term frequency-inverse document frequency (TF-IDF) (Rajaraman and Ullman, 2011) normalisation, and with and without pre-processing the training and test sets.

We then chose the two best models to compose the ensemble to be used at the competition. As it will be shown in the next section, these were Random Forests and Naïve Bayes. In the sequence, we also added an SVM classifier, to RBF kernel and $C = 2$ penalty to the ensemble, making Logistic Regression our meta-classifier.

The training set was divided into 90% for training/validation and 10% for test set. Models were trained in the training/validation set using 10-fold cross-validation. (Han et al., 2011).

[4]https://scikit-learn.org/stable/

6 Results

Tables 2 and 3 show the performance and settings of each classifier in the training/validation and test sets, respectively. During training, best results were observed without preprocessing, for RF and LR, whereas NB showed better results with preprocessing. These results, however, were very close to each other, ranging from $F1 = 0.69$ to $F1 = 0.71$. Regarding language model, best results were observed with 5-grams, for RF and LR, and 4-grams, for LR and NB.

At the test set, best results, for all methods, were observed with preprocessing the data. Normalising the vectors does not seem, however, to have influenced results when preprocessing is used. All best values were obtained with 4-grams. Overall, the best result was achieved with Naïve Bayes ($F = 0.74$), with preprocessing, using a 4-gram language model, and both with and without TF-IDF normalisation.

The ensemble model was tested with only one configuration: 4-Gram, with normalization, and without preprocessing. This configuration resulted in an $F1 = 0.729$ in the training set (a 2.5% increase over the best model in this set) and an $F1 = 0.751$ in the test set, corresponding to a 1.5% improvement over the best model in this set. As it turns out, especially in the test set, differences between the ensemble and its best constituent method do not seem so high.

7 Discussion

The competition rules allow only two models to be sent by each team. Although our Naïve Bayes model has shown good performance in the test set we had at hand, we chose not to send it to HaSpeeDe 2 due to the fact that it would also be tested in an out-of-domain data set.

Since this classifier can be very sensitive to domain changes, specially regarding null frequency words, which might bring the whole model down to multiplying smoothing values, we thought we would be better off not sending it. Still, it remained as one of the weak classifiers in the Ensemble we sent, so it was not completely put aside.

The organization of the competition presented F1 results corresponding to two classifiers, run in the same data set distributed to all participants in the competition. These were supposed to be taken as baselines by all competing teams. The first consisted of a majority class classifiers (Baseline-MC), which always chooses the majority class to label new examples. The second classifier, in turn, consisted of an SVM with linear kernel, running with TF-IDF normalisation (Baseline-SVM).

Table 4 shows the result of these two baseline classifiers, along with the classifiers we submitted to the competition (*i.e.* our Ensemble model and its constituent Logistic Regression classifier). As it turns out, for the within-domain task, only our Ensemble was superior to the baselines (3.9% over the baseline SVM and almost 123% over the majority class baseline). When moving to the out-of-domain test set, this difference dropped to only 1.8% over the SVM model and 62.3% over the majority class, still outscoring both baselines.

Regarding our Logistic Regression model, when run in the within-domain test set, it outscored only the majority class baseline (109% better), being however outscored by the baseline SVM by 2.3%. As for the out-of-domain test set,

Table 4: Result of baselines and final performance of classifiers in task A in terms of F1

Classifier	**Out-of-domain**	**In-domain**
Baseline-MC	0.3894	0.3366
Baseline-SVM	0.621	0.7212
Ensemble	0.632	0.749
LR	0.621	0.705

our Logistic Regression model presented the same result as the baseline SVM, outscoring the majority class baseline by 59.5%. Interestingly, both Ensemble and Logistic Regression models scored similarly in this set.

8 Conclusion

In this article we reported on the results obtained by two models submitted to EVALITA's HaSpeeDe2 task. Even though our Ensemble model outscored both benchmarks, we believe it could do better, should other choices regarding the language model be made.

Since the best results were obtained with longer word sequences (in our case, 4-grams), it might be the case that other language models, such as Glove or CBOW, for example, which make use of context words at both sides of the target word, could come up as better alternatives for the 4-gram model we used. BERT could also be a possibility to test.

Our best results were also obtained, at least during test, with preprocessing the data. We thus believe this is something to be kept. Regarding the normalisation of feature vectors, we could not observe great differences between using it or not, at least when it comes to TF-IDF normalisation.

Another direction to be followed might be to test other models as weak classifiers in the Ensemble, or even ensemble strategies other than stacking. This is something we leave for future work.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, USA, June.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA'18)*.

Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.