Vigilante at SemEval-2019 Task 6
How Far Can Models Overcome Data Imbalance

Ioan Budea
Imperial College London
ib2215@ic.ac.uk

Adrian Catană
Imperial College London
ac7815@ic.ac.uk

Abstract

Social media content generation is driven by the constant user growth. With this trend, the amount of hate speech is also increasing, making the necessity of automated methods for offensive language detection more proeminent. This paper presents our empirical results for the OffensEval challenge on identifying and categorizing offensive language in social media. We present different approaches, the different architectures of the models and discuss the results obtained and challenges encountered in designing classifiers for this purpose.

1 Introduction

With the emergence of social networking websites, the information posted and shared by users reaches a tremendous audience instantly. However, these websites and technologies have also enabled individuals to disseminate aggressive and harmful content as well, shielded by this layer of indirection. Finding a way to limit the generation and impact of offensive speech has attracted a lot of attention from the research community in recent years [Davidson et al., 2017, Malmasi and Zampieri, 2017]. The amount of new data posted every day makes manual monitoring and moderation of content infeasible.

In this paper, we present out findings on the SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media. The competition required building classifiers for three different subtasks: offensive language identification (Task A - binary classifier to detect offensive and acceptable language), automatic categorization of offense types (Task B - binary classifier to split the offensive content in Task A into targeted and untargeted) and offense target Identification (Task C - ternary classifier for targeted offensive language). The participants were provided with a dataset of 13200 labelled tweets to be used as training data.

In Natural Language Processing (NLP), classical methods required manual feature engineering and used statistical algorithms to perform the majority of the tasks. The new wave of NLP models borrow ideas from deep learning to perform extraction of hierarchical features from the given data. For our submissions, we decided to use Random Forests as an algorithm representative for the first category, and Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) for the Deep Learning approaches.

The code for our models and experiments is available at: https://gitlab.doc.ic.ac.uk/ac7815/nlp_cw.

2 Data Processing

The given dataset from the competition starting kit contains 13240 tweets: 4400 of them are marked as offensive and 8840 are not offensive. The offensive ones are further split for task B into 524 untargeted and 3876 targeted. Finally the targeted ones are divided for task C into 2407 targeted on individuals (IND), 1074 targeted on groups (GRP) and 395 targeted on others (OTH).

We have used the holdout method to split these tweets into a training set and a validation set with a ratio of 75:25. However, after carefully reviewing different parameters and results, the final submissions, that achieved our best scores in the competition, were trained on the entire dataset. As a small test set we have parsed an additional file present in the starting kit to construct a dataset of 320 tweets.

After we report the results of our validation experiments, we train the classifiers on the entire dataset before submission.

As a small test set we have parsed an additional file present in the competition starting kit to con-

struct a dataset of 320 tweets.

We convert all tweets to lowercase and remove the following elements as suggested in [Gaydhani et al., 2018]:

- unnecessary space patterns

- non-alphanumeric characters from the tweets

- remove the stopwords using the nltk Python package

- URLs present in the text

After applying data augmentation, we had a choice of using stemming, lemmatization (described in the subsections below) or none. Then, a vocabulary is built from all the words in the training set. Each word in the sentences is transformed into a number based on the index in this vocabulary. Then, each sentence is padded to the maximum length of all sentences by filling the vectors with the element 0 which represents the special padding symbol.

2.1 Data augmentation

From the above ratios, we see that the ratios of examples from each class are not comparable. While the issue is not so proeminent for the training dataset used for task A, this class imbalance might become problematic when investigating targeted offensive language. Therefore, we have decided to apply the following techniques for data augmentation:

- increasing the number of offensive samples by taking an offensive tweet and appending a random harmless tweet to create a new train instance

- increasing the number of untargeted samples by taking an untargeted tweet and duplicating it as another train instance

- increasing the number of samples labelled as 'other' ('OTH') by taking every such tweet and duplicating it as another train instance

Note that this augmentation process is applied twice, after the training-validation split is run, on both training and validation sets. The reason for this is that we do not want to duplicate tweets and risk to have them in both the actual training and validation sets - otherwise, the validation loss computed at each epoch would not give us trully "unseen" examples.

2.2 Stemming

We have also attempted word stemming for our tasks. Stemming can be defined as a rather crude heuristic process that removes the affixes of words in the hope of bringing the inflectional forms and derivationally related forms of a word to a common base form. In our work we have used the Porter stemming algorithm implemented in the same nltk Python package.

2.3 Lemmatization

While the goal of lemmatization is similar to the case of stemming, the latter method obtains the 'stem' of a word after applying a set of rules without taking into account the part of speech or the context of the word occurrence. Lemmatization obtains the base form of a word known as a 'lemma' of a word which involves reducing the word forms to its root form after doing a full morphological analysis of a sentence. While this technique may work in some application, it can hurt the performance of others. We have attempted this alternative pre-processing because stemming usually manages to increase recall while harming the precision of the classifiers, so a more informed decision to transform words may provide a better result. For this work, we have used a lemmatizer based on the WordNet lexical database of English words.

2.4 Slang

Considering the given datasets, the tweets that have been used might point to different informal expressions. An example of Internet slang is 'lol', which means 'laugh out loud', or AFAIK, meaning 'as far as I know'. Thus, this has been taken into account by using a Slang Dictionary, containing a list of 70 such expressions that are widely used - during the tweet parsing stage, the short-form slang is expanded to its original meaning.

3 Classifiers

Other work on hate speech in tweets [Badjatiya et al., 2017, Pitsilis et al., 2018] has been conducted using deep learning techniques. Our experiments use two of the methods proposed: Convolutional Neural Networks and Long short-term memory, comparing them against the general framework of Random Forests. Moreover, they concluded that embeddings learned from Deep Neural Network models when combined with gra-

dient boosted decision trees led to best accuracy values, which significantly outperforms the existing methods. In this section we describe the classifiers tested, the framework used for the implementation and the hyperparameters that can be tuned for each approach.

3.1 Random Forest

Random Decision Forests [Breiman, 2001] are an ensemble learning technique that constructs multiple decision trees using the training data and looks at the mode of the classes to decide the prediction when classifying new data. The parameters that can be modified are the number of estimators (i.e. the number of trees in the forest), the maximum depth of each tree (having a limit is a good technique for preventing overfitting) and the criterion for each spilt (a choice of either the entropy or gini measure). Using the RandomForestClassifier from the sklearn Python package, we have built random forests of 30 trees with a maximum depth of 80 nodes for each of the three tasks (the parameters have been chosen using grid search).

3.2 Convolutional Neural Network

A Convolutional Neural Network (CNN) [LeCun et al., 1998] is a feed-forward neural network consisting of the typical input and output layers, but with the hidden layers having additional types: convolution and pooling. The idea of the convolutional layers is to enable the network to automatically identify and extract features from the input space, while pooling layers can be introduced to reduce the dimentionality of the problem space. CNNs are mostly used in applications such as image and sound recognition, but they can be suited to particular NLP tasks as well. The intuition is that convolutional filters can correspond to several different semantic patterns [Jacovi et al., 2018] and thus, in our context of offensive language classification, CNNs can extract particular word combinations.

Our CNN is implemented using the PyTorch framework [PyTorch, 2019]. Once the input vectors are built using the index of words in the known vocabulary, these will be used as input to our CNN. The first layer is an embedding layer that transforms each word into a continuous vector space of given dimension [Mikolov et al., 2013]. Thus, the word embeddings will be contextualised and learned from the given examples. Next, the output of this embedding layer is fed into a con-

volutional layer, followed by a ReLU activation function, a max pooling layer and a dropout layer. At the end, a linear layer is applied to bring the result to either a single output value which will be transformed into a probability value for the binary classification task or a vector representing the corresponding outputs of all the classes (i.e. three classes for task C).

The hyperparameters that can be modified are: the number of layers, the dimension of the convolution kernels, the dropout probability, the optimization algorithm and learning rate used. We have opted for a shallow network architecture because the tweets have a small maximum length, so the number of dimensions does not need to be reduced dramatically. By varying the window size from the text taken into account in our convolutional layer we found that 1 yields the best results. The CNN was trained using the Adam optimizer [Kingma and Ba, 2014] to accelerate convergence during training with a learning rate of 0.0008 found using grid search.

3.3 Long-Short Term Memory

Recurrent Neural Networks (RNN) have been shown to produce very strong results for language modeling [Mikolov et al., 2010]. They are suited for language tasks because they take into account the whole input sequence, storing the previous results in a hidden state within the nework architecture. Long-short term memory [Hochreiter and Schmidhuber, 1997] is a very popular type of RNNs, proposed to address the problem of vanishing and exploding gradients and make RNN training feasible.

We have implemented an LSTM network in PyTorch, using an embedding layer, a layer of LSTM cells and a fully connected layer for producing the output. We have opted for a unidirectional LSTM architecture because of the small maximum length of the tweets, coupled with the small number of training examples for a network with so many learnable parameters. The LSTM was trained using the same parameters for the optimization procedure as our CNN: Adam optimizer with the same learning rate.

4 Results

This section presents the best runs for the models described in section 3, for each individual task.

4.1 Validation Experiments

The results of our validation experiments are summarized per task in Tables 1, 2 and 3.

For task A, our CNN was trained for 100 epochs, while the LSTM had a tendency to diverge, so we report the best results computed at any point during training as we would have to apply early stopping to obtain a reasonable model.

What we can observe is that the accuracy results for the Random Forest are not very far from the actual ratio between the training examples. This may indicate that because of the data imbalance, the model will tend to predict only one label, classifying all test instances as not offensive since this is the majority class.

On the other hand, the CNN trained using stemming manages to produce the best results in our validation experiments both in terms of accuracy and F1-measure. When we applied data augmentation for this task as explained in section 2.1, the F1-measure increased up to 0.52 at the expense of the test accuracy which dropped to a value of 67.19% on the best run.

For task B, we have decided to reduce the number of training epochs to 80 since there were less examples remaining (only the subset of offensive tweets need further classification).

For task C, the performance of Random Forest was considerably inferior to the other techniques, so we are not including it in our comparison. One explanation for this could be the fact that such a classifier will make decisions based on individual words, treating all of them as independent. For a complicated task where multiple levels of the language and the context need to be taken into account, this independence assumption and local decision at each step leads to a limited generalization.

Using data augmentation for training our CNN, together with stemming, we have improved the result up to a test accuracy of 71.79% and F1-measure of 0.21, considerably better than any results on the original, unmodified dataset.

4.2 Submission Results

For the actual contest submissions, we can view only the best result for each task. We report the best macro-averaged F1-score:

1. Task A: a result of **0.658273895** - this was obtained for the CNN trained on the full training dataset pre-processed using the common techniques from section 2 and stemming

2. Task B: a result of **0.648093842** - obtained again for the CNN, but trained on the **augmented** training dataset pre-processed using the common techniques from section 2 and stemming

3. Task C: a result of **0.400910364** - CNN obtained directly from the validation experiments without further tuning

4.3 Discussion and Interpretation

The results show consistently that using stemming instead of no further processing or lemmatization was a better choice for these tasks. This can be motivated by the fact that the language used may not be very complex or have many connotations. In that case, a simpler heuristic would yield reasonable results.

Another point to note is that our vocabulary is built using the words that we see in the train corpus. This intrinsically assumes that they are representative for the words that will be included in the test set. If this does not hold, the method does not provide the best way to handle unseen words since these will simply be discarded and will not contribute to the classification. However, one should also investigate the possibility of obtaining better results if the whole English vocabulary or at least a bigger subset of it is used for obtained word indices, something we have not experimented with in our work.

5 Conclusion

In this paper we have summarized the approaches and methods used for building simple classifiers for offensive speech detection. We wanted to analyze how the class distribution over the dataset affects the generalization capabilities of our algorithms, and we have seen that this was not a problem up to a certain level of class imbalance, beyond which all the methods find that it is rewarding to stick to the majority class. Applying data augmentation techniques for the extreme cases of imbalance such as the samples for task B proved to be the only viable option. In general, the problem of obtaining accurate data that resembles the true underlying probability distribution is the challenging aspect of every deep learning method for natural language processing.

Task A	baseline	stemming	lemmatization
Random Forest	Test Acc: 72.81%	Test Acc: 70.31%	Test Acc: 75.31%
CNN	Test Acc: 82.50% F1-measure: 0.49	Test Acc: 85.94% F1-measure: 0.65	Test Acc: 81.25% F1-measure: 0.41
LSTM	bold	bold	bold

Table 1: Task A experimentation results.

Task B	baseline	stemming	lemmatization
Random Forest	Test Acc: 53.24%	Test Acc: 51.94%	Test Acc: 53.24%
CNN	Test Acc: 68.83% F1-measure: 0.74	Test Acc: 58.44% F1-measure: 0.67	Test Acc: 63.64% F-measure: 0.70
LSTM	bold	bold	bold

Table 2: Task B experimentation results.

Task C	baseline	stemming	lemmatization
CNN	Test Acc: 71.79% F1-measure: 0.12	Test Acc: 69.23% F1-measure: 0.11	Test Acc: 69.23% F1-measure: 0.11
LSTM	Test Acc: 66.67% F1-measure: 0.10	Test Acc: 66.67% F1-measure: 0.37	bold

Table 3: Task C experimentation results.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv e-prints*, art. arXiv:1703.04009, Mar 2017.

Shervin Malmasi and Marcos Zampieri. Detecting Hate Speech in Social Media. *arXiv e-prints*, art. arXiv:1712.06427, Dec 2017.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and TFIDF based approach. *CoRR*, abs/1809.08651, 2018. URL http://arxiv.org/abs/1809.08651.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets. *arXiv e-prints*, art. arXiv:1706.00188, Jun 2017.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. Detecting Offensive Language in Tweets Using Deep Learning. *arXiv e-prints*, art. arXiv:1801.04433, Jan 2018.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1): 5–32, October 2001. ISSN 0885-6125. doi: 10. 1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proc. of IEEE*, 1998.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding Convolutional Neural Networks for Text Classification. *arXiv e-prints*, art. arXiv:1809.08037, Sep 2018.

PyTorch, 2019. URL https://pytorch.org/.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, art. arXiv:1301.3781, Jan 2013.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

T. Mikolov, M. Karafiat, L. Burget, J. Cernock, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of Interspeech*, 2010.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.