AdelaideCyC at SemEval-2020 Task 12: Ensemble of Classifiers for Offensive Language Detection in Social Media

Mahen Herath[1], Thushari Atapattu[2], Hoang Anh Dung[2], Christoph Treude[2], Katrina Falkner[2]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
mahenherath@gmail.com, dunganh.hoang@student.adelaide.edu.au,
{thushari.atapattu, christoph.treude, katrina.falkner}@adelaide.edu.au

Abstract

This paper describes the systems our team (AdelaideCyC) has developed for SemEval Task 12 (OffensEval 2020) to detect offensive language in social media. The challenge focuses on three subtasks – offensive language identification (subtask A), offense type identification (subtask B), and offense target identification (subtask C). Our team has participated in all the three subtasks. We have developed machine learning and deep learning-based ensembles of models. We have achieved F1-scores of 0.906, 0.552, and 0.623 in subtask A, B, and C respectively. While our performance scores are promising for subtask A, the results demonstrate that subtask B and C still remain challenging to classify.

1 Introduction

The surge of Internet and social media technologies provides a wealth of opportunities for cybercrime, and has led to the unprecedented social crisis of online abuse. Despite the illegality of such behaviour, most social media platforms such as Facebook, Twitter, Instagram are susceptible to online bullying due to their openness and anonymisation. The sheer amount of offensive language generation vastly exceeds the capacity of manual detection. Therefore, there is a crucial need for urgent development of technological solutions. The automated identification of offensive language has been recognised as a subtask of NLP only recently and most of the advances have occurred in the last few years.

The offensive language identification as an NLP problem is inherently complex and challenging – even for humans (aside from the offensive language's victims) due to many variants of language used by harassers such as coarse language, sarcasm, intimidation, and colloquialisms. People also tend to use coarse language in a friendly manner, without an intention to harm anyone. Therefore, it is important to identify whether a post or a tweet is offensive and whether it is targeted at an individual or a group. In this paper, we focus on identifying offensive posts extracted from the Twitter platform. The training dataset contained more than 9 million tweets and they were annotated using a semi-supervised approach. Our team has participated in English versions of all three subtasks organised by Zampieri et al. (2020), i.e., subtask A: offensive language identification (offensive or not), subtask B: offense type identification (targeted insults and threats or untargeted), and subtask C: offense target identification (individual, group, other). In this paper, we discuss the models developed for each subtask along with the performance. The advancement of offensive language identification has many benefits for social media and online communities to protect their users.

2 Related work

The identification of unacceptable language in social media and online communities has attracted attention from researchers in related fields such as cyberbullying (Rosa et al., 2018), aggression (Kumar et al., 2018), hate speech (Fortuna and Nunes, 2018), abusive language (Waseem et al., 2017),

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1516–1523
Barcelona, Spain (Online), December 12, 2020.

and offensive language (Zampieri et al., 2019). Davidson et al. (2017) is one of the first studies to create a dataset for offensive language detection by categorising tweets into hate speech, offensive but not hate speech, and neither. Their work utilised various features such as n-grams, TF-IDF, readability scores, and sentiments to build machine learning models like logistic regression and SVM. Recently, multiple classification tasks like OffensEval (Zampieri et al., 2019) and HatEval (Basile et al., 2019) contributed to the advancement of the research field by creating datasets and tasks to identify offensive language, type, and target (Zampieri et al., 2019) and hate speech, target, and aggressiveness (Basile et al., 2019).

The systems developed for these tasks used cutting-edge NLP, machine learning and deep learning techniques. Some key systems for OffensEval and HatEval such as Fermi (Indurthi et al., 2019) used Universal Sentence Encoder to build a SVM model, NLPR@SAPOL (Seganti et al., 2019) used an ensemble of deep learning models like OpenAI Finetune and Transformer, while NULI (Liu et al., 2019) developed a BERT-based model. Although some systems have achieved reasonable performance (e.g., 0.82 F1-score for subtask A of OffensEval by NULI), most other systems still lack 'good' performance for other subtasks such as identifying target and type of offenses. Some of these challenges focus on specific problems like hate speech against minorities (women, migrants) (Basile et al., 2019) while OffensEval classification tasks focus on 'general' offensive language available on Twitter.

3 Methods

3.1 Dataset

Table 1 includes the data description of the training dataset. Instead of labels, the training dataset provided by the organisers includes average confidence values. For subtask A, we have considered 0.5 as the threshold and categorised tweets as 'offensive' when the average confidence is greater than 0.5 and 'not-offensive' otherwise. Similarly, in subtask B, average confidence greater than 0.5 is considered as 'targeted' and 'untargeted' otherwise. For subtask C, we considered maximum average confidence as the measure to determine 'individual (IND)', 'group (GRP)', and 'other (OTH)' labels. The test dataset included 3887, 1422, and 850 tweets for subtasks A, B, and C respectively. More information about the dataset and the annotation process is included in Zampieri et al. (2020) and Rosenthal et al. (2020).

Subtask	**Offensive**				**Not-offensive**	**Total**
A	1,449,656				7,637,462	9,087,118
B	Targeted			Untargeted		
	149, 550			39,424		188,974
C	Individual	Group	Other			
	93,638	16,768	3,396			113,802

Table 1: Data description of training dataset.

3.2 Preprocessing

The datasets for all three subtasks have been sourced from Twitter. Thus slang words, abbreviations, misspelled words and emoticons etc. are abundant in data instances. Therefore we carried out a few pre-processing steps to clean the datasets. These steps include replacing slang words and abbreviations[1], decoding emoticons[2] and removing non-ascii characters from the dataset. In addition to this, several standard data pre-processing steps such as removal of punctuation and URLs were inherently performed while fine-tuning deep learning based language models like DistilBERT (Sanh et al., 2019).

[1] https://floatcode.wordpress.com/2015/11/28/internet-slang-dataset/

[2] https://github.com/carpedm20/emoji

3.3 Data preparation

A significant class imbalance was observed in the training datasets of all three subtasks. In subtask A, a binary classification problem, 84.05% of the tweets in the training dataset belonged to the class 'NOT' and only 15.95% of tweets belonged to the class 'OFF'. In subtask B, a binary classification problem, 78.4% of the tweets in the training dataset were labeled 'TIN' while only 21.6% of the tweets were labeled 'UNT'. In subtask C, a multi-class classification problem, 82.28% of the tweets in the training dataset belonged to class 'IND', with only 14.73% and 2.98% of the tweets belonging to classes 'GRP' and 'OTH' respectively. To mitigate adverse effects of class imbalance, we experimented with downsampling the majority class instances in the training datasets for subtask A and B. Similarly in subtask C, where we employed a one-vs-all strategy to train binary classifiers, we downsampled the majority class instances accordingly.

4 Models and Results

4.1 Subtask A

We used DistilBERT (Sanh et al., 2019), a lighter, faster version of BERT (Devlin et al., 2019), to create four classification models A, B, C and D for subtask A. Model A was trained on a downsampled and balanced subset of training data while models B and C were trained on imbalanced subsets of training data where the majority classes were 'OFF' and 'NOT' respectively. Drawing inspiration from Khoussainov et al. (2005), model D was trained on a balanced subset of the training data composed of tweets which were assigned opposing class labels by the two biased classifiers B and C. All three models were finetuned with a learning rate of 5e-5 for 2 epochs using a batch size of 32.

We then created an ensemble classifier combining the models B, C and D using a voting scheme. If the two biased classifiers B and C agreed upon a predicted label, the data instance was assigned that particular label. In case they disagreed, we assigned the prediction made by model D. Thus model D served as a tie-breaker. We also created another ensemble classifier based on a majority voting scheme using models A, B and C. All our models for subtask A were trained and tested on the Google Colaboratory platform[3].

We evaluated the performance of our classifiers against three different distributions of held-out validation data. Dataset A was a balanced subset of validation data, while datasets B and C were imbalanced subsets of validation data with majority of 'OFF' and 'NOT' labels respectively. Table 2 shows the results of our experiments. Our official submission to the competition was made using the ensemble model B + C + D. Table 3 shows our performance in comparison with the competition results.

	Macro Averaged F1-Score		
Model	**Dataset A**	**Dataset B**	**Dataset C**
A	0.9468	0.9332	0.9096
B	0.9270	0.8640	0.9300
C	0.9244	0.9385	0.8498
A + B + C	0.9490	0.9329	0.9124
B + C + D	**0.9542**	**0.9403**	**0.9240**

Table 2: Performance of the models on the evaluation datasets of subtask A

[3] https://colab.research.google.com/

System	F1-Score
Top system	0.922
Our system	**0.906**
Baseline 1	0.419
Baseline 2	0.419

Table 3: The results of subtask A in comparison with the competition.

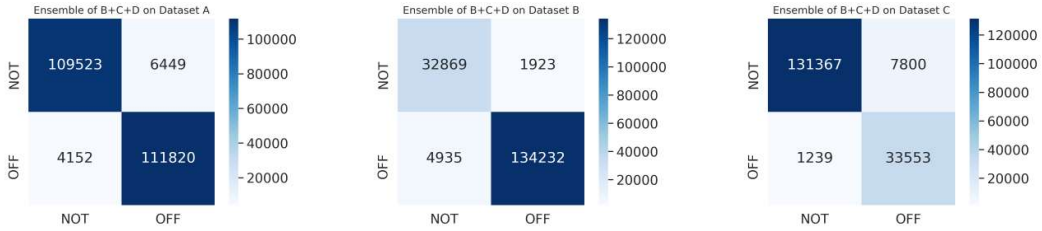

Figure 1: Confusion matrices of the best performing ensemble model for evaluation datasets A, B and C of subtask A

According to the results in Table 2, we have achieved a 0.95 macro-averaged F1-score for our combination of models B, C and D using the dataset A. All other datasets also showed promising performance with F1-score greater than 0.92. This robustness of the model is also evident from the confusion matrices shown in Figure 1. According to the results in Table 3, we achieved comparable results with the top system. The F1 difference is 0.016.

4.2 Subtask B

For subtask B, we experimented with machine learning models such as Logistic Regression, Linear SVC, and a neural network model - CNN-LSTM. We also fine-tuned transformer models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) since these pre-trained language models demonstrate state-of-the-art performance for downstream NLP tasks. To train Logistic Regression and Linear SVC, we used TF-IDF vectors as features. We used default hyperparameters for neural network and transformer models and the best performance was achieved with 3 epoches. Performance of these single classifiers was measured against a held-out evaluation dataset. We have achieved more than 0.87 F1-score with all experimented models while XLNet showed the best performance of 0.889. However, when single classifiers were further experimented with the test dataset from OffensEval 2019 (Zampieri, 2019), we experienced a drop in performance using single classifiers. Therefore, we experimented with ensemble models by averaging predictions from combinations of single classifiers to deduce the final predictions for the test dataset. Table 4 shows the best models with stable performances. We have selected the ensemble of Logistic Regression, LinearSVC, RoBERTa, XLNet and BERT as our most robust model across different distributions of testing data.

System	F1-Score
XLNet	0.889
CNN-LSTM	0.876
Ensemble	0.890

Table 4: The performance of the models on the evaluation dataset of subtask B .

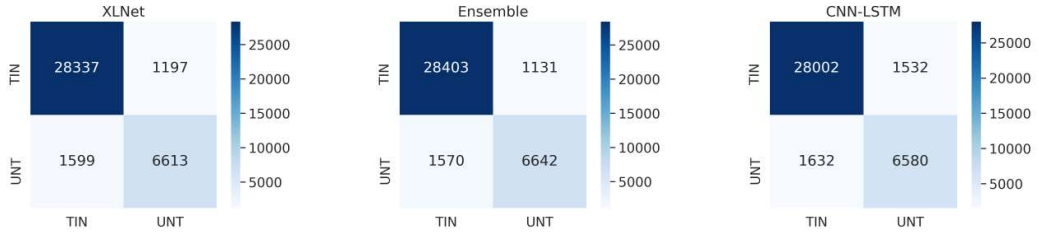

Figure 2: Confusion matrices of top performing models for subtask B

The test set of subtask B consisted of unlabelled 1,422 data points, each required to be predicted as either targeted insult and threat (TIN) or untargeted (UNT). Table 5 shows the performance of our ensemble model using the test dataset.

System	F1-Score
Top system	0.746
Our system	**0.552**
Baseline 1	0.374
Baseline 2	0.374

Table 5: The results for subtask B in comparison with the competition.

Even though our performance was good using the held-out set, we observed low performance (F1score of 0.55) of our system when applied to the test set. This drop could be occured due to the large class imbalance in the dataset (i.e. TIN class is approximately 4 times bigger than UNT class - see Table 1). We also observed a difference in the class distribution between the training dataset and the official, labelled test dataset. In the training dataset 78.4% of all tweets belong to the class 'TIN'. However in the official test dataset only 59.7% tweets belong to the same class. Similarly while only 21.6% of tweets in the training dataset are labelled as 'UNT', 40.2% of test tweets belong to the class 'UNT'. Since this is quite prevalent in many real world problems, the need to design more robust models is highlighted through these results. Further, a manual analysis of a sample of misclassified tweets suggested that our threshold of 0.5 to distinguish TIN and UNT classes is quite ambiguous in some instances.

4.3 Subtask C

We reduced the multi-class classification problem of subtask C into separate binary classification sub tasks. According to the problem description, every training data instance can belong to only one of the given three classes, 'IND', 'GRP' or 'OTH'. Therefore we first trained two binary classifiers, one to predict whether a given data instance belongs to the class 'IND' and the other to predict whether the given data instance belongs to the class 'GRP'. We finetuned each model with a learning rate of 2e-5 and a batch size of 32, for 2 epochs. We then combined the predictions from two classifiers to retrieve final class labels. If a data instance was marked as positive by either of the classifiers, we assigned the class label corresponding to that classifier. Whenever there was a tie, we selected the prediction with the highest probability score, while giving precedence to 'GRP' class when highest probability scores were equal. We assigned the label 'OTH' to instances which were marked as negative by both classifiers. Each classifier was trained using DistilBERT (Sanh et al., 2019) on balanced subsets of training data. Our official submission to subtask C was made using this model.

In addition, we trained a third binary classifier to distinguish instances belonging to the class 'OTH' using DistilBERT, and created an ensemble of the three classifiers. Whenever the positive predictions from a pair of classifiers or all three classifiers resulted in a tie, we selected the prediction with the highest probability score. Whenever all three classifiers predicted negative for a given data instance, we selected the prediction with the lowest probability score to break the tie.

This newer ensemble model was created after the official deadline for subtask C, and hence we could not submit it for the challenge. Yet, after the official, labelled test dataset of subtask C was made available at the end of the challenge, we evaluated our system and observed a macro averaged F1-score of 0.6719, which would have ranked 2nd amongst all submissions for subtask C.

Ensemble Model	Macro averaged F1-Score
IND + GRP	0.6351
IND + GRP + OTH	0.7064

Table 6: Performance of the ensemble models on the evaluation dataset of subtask C

System	F1-Score
Top system	0.714
Our system	**0.623**
Baseline	0.005

Table 7: The results of subtask C in comparison with the competition.

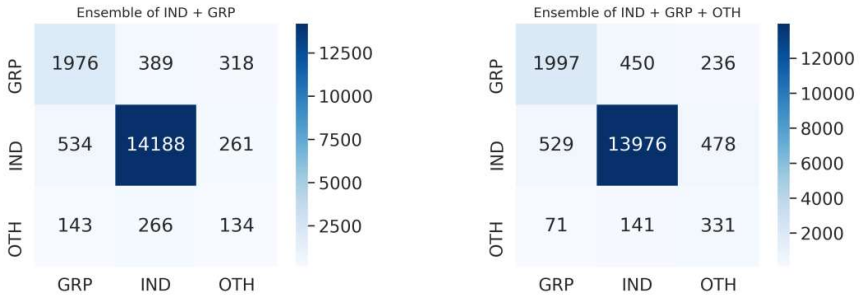

Figure 3: Confusion matrices of the two ensemble models for the evaluation dataset of subtask C

As evident from the confusion matrices in Figure 3, both ensemble models perform relatively well when identifying 'IND' and 'GRP' instances, but perform poorly when identifying 'OTH' instances. When experimenting on the held-out evaluation dataset, single classifiers trained to identify 'IND' and 'GRP' instances reported F1 scores of 0.8765 and 0.8648 respectively, while the single classifier for 'OTH' instances reported an F1-score of 0.6071. This drop could be attributed to the scarcity of training instances belonging to class 'OTH'. While the first two single classifiers were trained on balanced samples having 25,810 and 21,462 data instances respectively, the classifier for 'OTH' class was trained on a balanced sample having just 4,348 data instances. Having more training data examples of class 'OTH' would have helped improve the performance of the latter classifier, and subsequently the overall performance of the ensemble model.

5 Conclusion

This paper presents the description of the systems we developed for SemEval 2020 Task 12. For subtask A and C, we have developed ensembles of models using DistilBERT. In subtask B, our best performing model was an ensemble developed using Logistic Regression, LinearSVC, RoBERT, XLNet and BERT. We have achieved promising results for subtask A relative to other systems in the competition. Despite the good results we have obtained for subtask B and C using the held-out set, our systems could be further improved by optimizing hyperparameters for subtask B and C and by

experimenting with various other features such as personal mentions, named entities etc., particularly for machine learning models in subtask 2.

References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, In the Proceedings of the 13th International Workshop on Semantic Evaluation.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In the Proceedings of the NAACL 2019, pp. 4171–4186.

Fortuna, P., & Nunes, S., 2018. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), 1-30. doi:10.1145/3232676.

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., & Varma, V. 2019. FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. In the Proceedings of the 13th International Workshop on Semantic Evaluation.

Khoussainov, R., Heß, A. and Kushmerick, N., 2005. Ensembles of biased classifiers. In the Proceedings of the 22nd international conference on Machine learning - ICML.

Kumar, R., Ojha, A. K., Malmasi, S. and Zampieri, M., 2018. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC).

Liu, P., Li, W., & Zou, L., 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In the Proceedings of the 13th International Workshop on Semantic Evaluation.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., Trancoso, I., 2019. Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345. doi:https://doi.org/10.1016/j.chb.2018.12.021.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., Nakov, P., 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. arXiv preprint

Sanh, V. , Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108

Seganti, A., Sobol, H., Orlova, I., Kim, H., Staniszewski, J. Krumholc, T. and Koziel, K., 2019. NLPR@SRPOL at SemEval-2019 Task 6 and Task5: Linguistically enhanced deep learning offensive sentence classifier. In Proceedings of The 13th International Workshop on Semantic Evaluation.

Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In the Proceedings of the First Workshop on Abusive Language Online.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le., Q. V., 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in Neural Information Processing Systems (NeurIPS).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In the Proceedings of the 13th International Workshop on Semantic Evaluation.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç., 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In the Proceedings of Semantic Evaluation workshop 2020.