Automated Detection of Cyberbullying Against Women and Immigrants and Cross-domain Adaptability

Thushari Atapattu[1], Mahen Herath[2],Georgia Zhang[1], Katrina Falkner[1]
[1]School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
[2]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Abstract

- Due to many negative impacts of cyberbullying, it is of crucial importance to detect abusive content published in social media platforms
- In this work, we use a Twitter dataset on hate speech against women and immigrants from SemEval 2019 challenge (Task 5) and create ensembles of models to detect offensive Tweets and to determine the targeted groups
- We analyze the misclassified Tweets using the open coding technique
- We also evaluate the cross-domain adaptability of the developed models

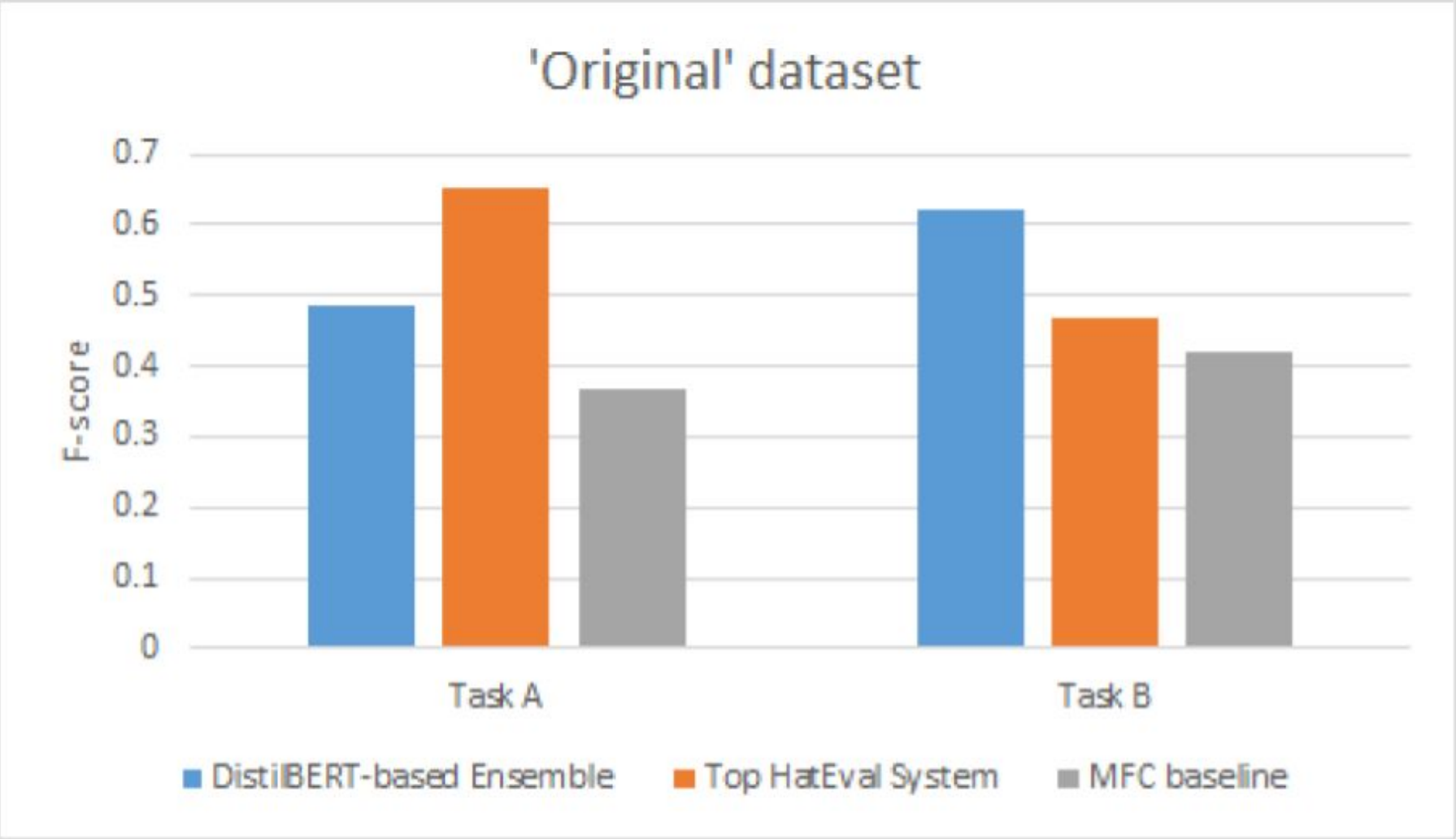
Image source : Ilayza Macayan on Unsplash

Dataset

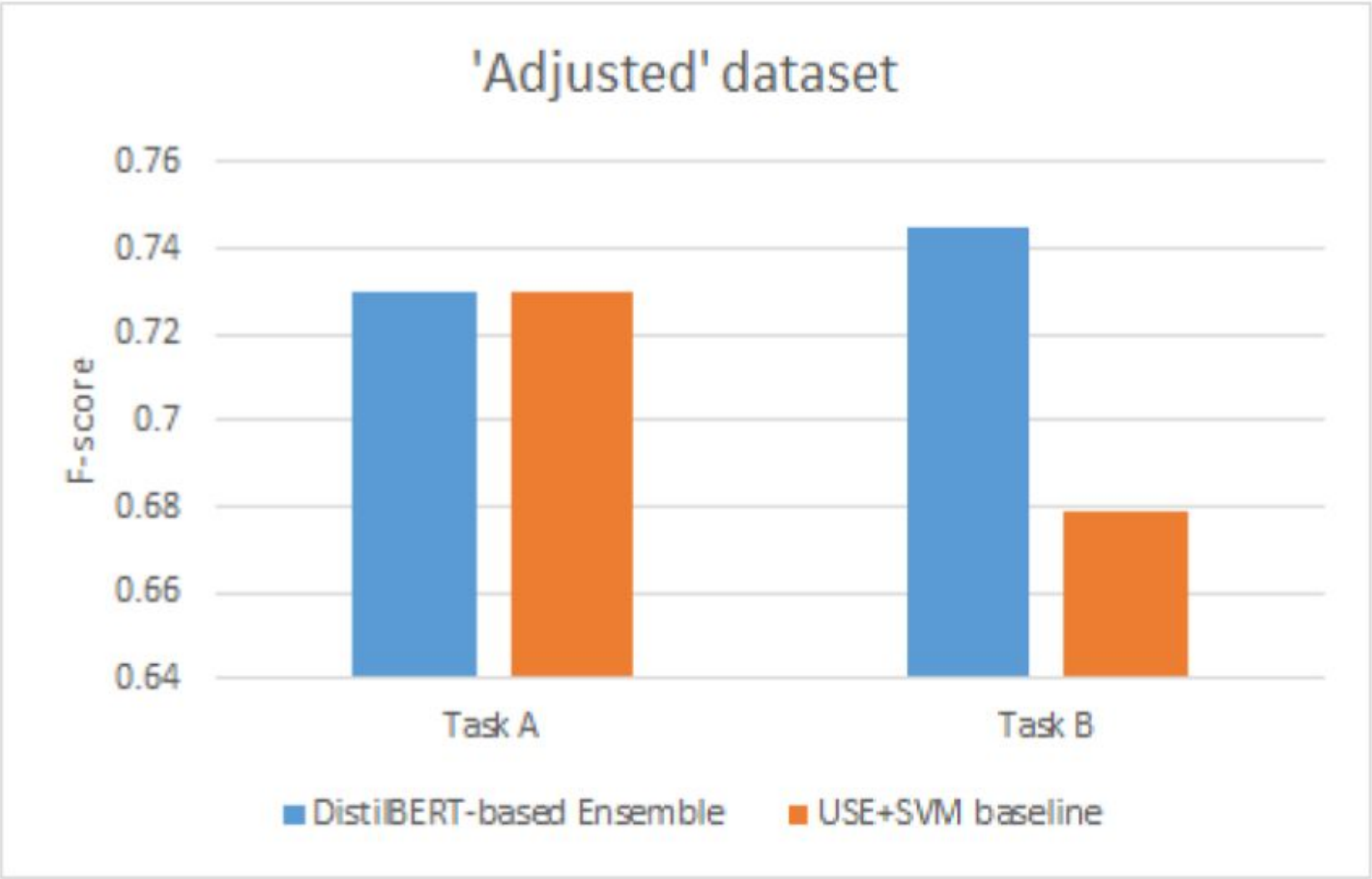
- We utilize the Twitter dataset (English) made available for SemEval 2019-Task 5 (Detection of Hate Speech Against Women and Immigrants in Twitter) Challenge [1].
- The dataset contains a set of tweets and their labels; HS - Hate Speech (0 - No, 1 - Yes), TR - Target Range (0 - generic group, 1 - individual), AG - Aggressiveness (0 - No, 1 - Yes)
- Training dataset contains 10000 tweets while the test dataset contains 3000 tweets.

Models

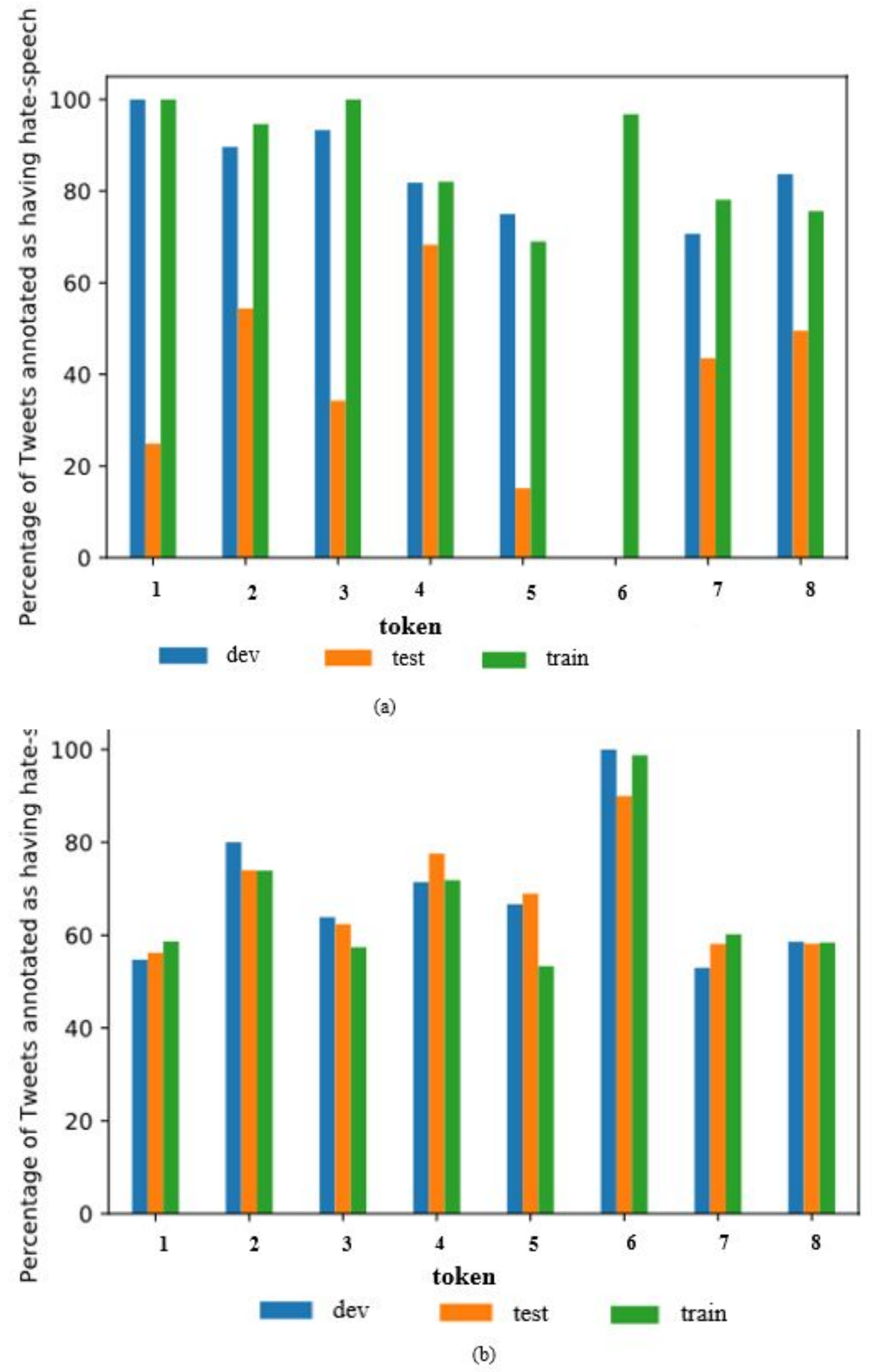
- We built ensembles of models based on DistilBERT, a lighter, faster version of BERT to tackle,
  - Task A - Detecting Tweets with content offensive to women and/or immigrants
  - Task B (Predicting the targets of offensive content)



Performance of the models for Task A & B while using 'original' HatEval test dataset



Performance of the models for Task A & B while using 'adjusted' HatEval test dataset



Variation of data across training, dev and test sets in (a) 'original' (b) 'adjusted' dataset

1:#buildthatwall, 2:#buildthewall, 3:#nodaca, 4:#sendthemback, 5:#stoptheinvasion, 6:#womens**k, 7:b***h, 8:h*e

Content Analysis of Misclassified Tweets

Code	Example	Explanation	%
GEND	*You seem like a h*e Ok b***h?Did I ever deny that? Nope, Next.*	Misogynistic if 'you' refers to a female	11
CNTX	*@user you deserve alllll the hate*	Misogynistic if '@user' refers to a female	44
SLNG	you a gay a*s b***h who seeks attention, STOP! I knew ever since you gonna switch up on me... I guess you did F***ING SNAKE A*S H*E!	Non-misogynistic if 'gay a*s b***h' slang is resolved	9
ERROR	*@user It means<religion> will show them how to rape/abuse women 24/7!*	This is targeted at immigrants, but has been erroneously annotated	3
MSCL	*Europe is being invaded by third world "refugees" Continue to Pray for them*	This is targeted at immigrants, but was misclassified by our model	3
OTHER	*REFUGEES NOT WELCOME*	This is targeted at refugees, but is incorrectly classified as negation has not been recognized	30

GEND - Gender related, CNTX - Lack of context, SLNG - Issues in resolving the slang, ERROR - Issue with original annotation, MSCL - Misclassified by our model, OTHER - Does not belong to any other category

Cross-task Evaluation

Dataset	Sample Size	Accuracy	Precision	Recall	F1 Score
OffensEval2020 [2]	3887	0.74	0.72	0.74	0.68
OffensEval2019 [3]	860	0.68	0.66	0.68	0.67
Hate & Offense [4]	2971	0.70	0.74	0.70	0.69

Discussion

- Discrepancies between the training and test data have an impact on the performance of the models.
- Misclassified tweets can be categorised into six types, with the context-related issues ('CNTX') being the most frequent reason for misclassification, followed by issues related to resolve gender ('GEND') and slang ('SLNG').
- Our pre-trained models are able to detect hate speech in other benchmarking datasets with a reasonable accuracy (~0.7).

References

[1] V. Basile et al., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter", *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019. Available: 10.18653/v1/s19-2007

[2] 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.

[3] Zampieri, M., Malmasi, S., Nakov, P.,Rosenthal, S., Farra, N., and Kumar, R.,2019.SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensE-val). In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 75–86.

[4] Davidson, T., Warmsley, D., Macy, M.,Weber, I.,2017. Automated hate speech detection and the problem of offensive language. In ICWSM.