

Contents

1	Introduction	3
1.1	Problem Description	3
1.2	Abusive Language and Hate Speech	5
1.2.1	Abusive Language Detection	6
1.3	Offensive Language	7
1.4	Abusive Language Detection in Low-Resource Settings	8
1.5	Motivation and Objectives	9
1.6	Research Questions	11
1.7	Contributions of this Thesis	12
1.8	Structure of the Thesis	12
I	Keyword Extraction and Bias Analysis	17
2	Offensive Keyword Extraction based on BERT	19
2.1	Introduction	20
2.2	Related Work	22
2.2.1	Automatic Keyword Extraction.	22
2.2.2	Text Representation based on Graph.	23
2.3	The Problem	23
2.4	Keyword Extraction based on BERT	24
2.4.1	Attention from BERT	25
2.4.2	Graph Representation	28
2.4.3	Keyword Extraction from Graph	28
2.5	Experiments	29
2.5.1	Results	31
2.6	Discussion	34
2.6.1	Error Analysis	35
2.6.2	Limitations of Our Work	35
2.7	Conclusion and Future Work	36

3	Keyword and Bias Analyses in Hate Speech Detection	39
3.1	Introduction	40
3.2	Related Work	42
3.3	Datasets	44
3.4	Transformer-based Models for Hate Speech Detection: Analysis of Salient Words	45
3.5	Analysis of Hateful Keyword	48
3.5.1	Method for Automatic Keyword Extraction	49
3.5.2	Experimental Setup	51
3.5.3	Discussion.	52
3.6	Bias Mitigation	54
3.6.1	Experimental Setup	54
3.6.2	Results and Discussion	57
3.7	Limitations and Ethical Concerns	60
3.8	Conclusions	61
II	Graph-Based Exploration	63
4	Graph Auto-Encoders for Multi-Domain and Multilingual Hate Speech Detection	65
4.1	Introduction	66
4.2	Related Work	68
4.3	Graph Auto-Encoders for Hate Speech Detection	69
4.3.1	Formalization	69
4.3.2	Background: Graph Auto-Encoders	69
4.3.3	Auto-Encoder Architecture	71
4.4	Experimental Design	72
4.4.1	Dataset	72
4.4.2	Experimental Setup	73
4.5	Embeddings Evaluation	73
4.5.1	Analysis of Latent Representation	73
4.5.2	Evaluation for Hate Speech Detection	74
4.6	Multi-domain Evaluation	75
4.7	Multilingual Evaluation	78
4.8	Conclusion and Future Work	80
5	Convolutional Graph Neural Networks for Hate Speech Detection in Low-Resource Settings	83
5.1	Introduction	84
5.2	HaGNN Model	85
5.2.1	Hate Speech Detection	85
5.2.2	Background: Convolutional Graph Neural Networks	86
5.2.3	Our Model	86

5.2.4	Proposed Loss: Similarity Penalty	87
5.2.5	Training the Model	88
5.3	Experiments	88
5.4	Results	89
5.5	Conclusions and Future Work	90
III Data Augmentation		93
6	Vicinal Risk Minimization for Few-Shot Cross-lingual Transfer in Abusive Language Detection	95
6.1	Introduction	96
6.2	Background and Related Work	98
6.3	Dataset and Experimental Setup	99
6.4	Few-Shot Cross-lingual Transfer	100
6.4.1	SSMBA	101
6.4.2	MIXUP	101
6.4.3	MIXAG	102
6.4.4	Multilingual MIXUP/MIXAG	103
6.4.5	Multidomain MIXUP/MIXAG	103
6.4.6	Results and Analysis	103
6.4.7	Ablation Studies	106
6.5	Unsupervised Language Adaptation	107
6.5.1	Results and Analysis	108
6.6	Conclusions and Future Work	110
6.7	Limitations and Ethical Concerns	110
IV Summary		117
7	Discussion of the Results	119
7.1	Keyword Extraction and Bias Analysis	120
7.1.1	Experimental Setup	121
7.1.2	Analysis of Abusive Keywords	121
7.1.3	Bias and Performance Analysis	122
7.2	Graph-Based Exploration	124
7.2.1	Experimental Setup	125
7.2.2	Results and Discussion	126
7.3	Data Augmentation	130
7.3.1	Experimental Setup	130
7.3.2	Data Augmentation	131
7.3.3	Results and Discussion	131
7.4	Hate Speech Spreaders	132
7.4.1	Author Profiling Shared Task	133

7.4.2	Users Networks Analysis	134
7.5	Ethical Discussion	140
8	Conclusions and Future Work	141
8.1	Conclusions	141
8.2	Future Work	144
8.3	Publications	145
	Bibliography	147