

Table of Content

1	Introduction	1
1.1	Problem Statement	2
1.2	Hypothesis	2
1.3	Objectives	3
1.3.1	General Objective	3
1.3.2	Specific Objectives	3
1.4	Methodology	4
1.5	Thesis Structure	5
2	Background and Related Work	6
2.1	Scientific Disciplines	6
2.1.1	Artificial Intelligence (AI)	6
2.1.2	Machine Learning	7
2.1.3	Natural Language Processing (NLP)	10
2.2	Text Classification	12
2.2.1	Task Formalization	12
2.2.2	Multi Label Text Classification	13
2.2.3	Extreme Multi Label Text Classification (XMC)	14
2.3	Clinical coding	15
2.3.1	Clinical coding in languages other than English	16
2.3.2	Codiesp	16

2.3.3	Cantemist	17
2.3.4	FALP	17
3	Divide and Conquer - DaC	18
3.1	DaC Corpus Preprocessing	18
3.2	Matcher	20
3.2.1	Transformers	20
3.2.2	Settings Matcher	22
3.3	Ranker	23
3.3.1	TF-IDF	24
3.3.2	One Vs Rest	25
3.3.3	Gradient Boosting Trees	26
3.3.4	Settings XGBoost	27
3.4	Combining results of the Matcher and Ranker	28
3.5	Ensemble	29
3.6	Data Augmentation using Named Entities	30
3.7	Library	31
3.7.1	Library important classes	31
3.7.2	Library files	37
4	DaC on multiple medical corpora	39
4.1	Corpora	39
4.1.1	CodiEsp	40
4.1.2	Cantemist	40
4.1.3	FALP	41
4.2	Ontologies and Cluster Choice	42
4.2.1	Ontology - ICD-10-CM	42
4.2.2	Ontology - ICD-10-PCS	43

4.2.3	Ontology - ICD-O-3	44
4.3	Metrics	46
4.4	Evaluation	47
4.5	Results	48
4.6	Module Analysis	50
5	Conclusions and Future Work	55
5.1	Conclusions	55
5.2	Future Work	55
5.3	Contributions	56
	Bibliography	66