Contents

1 Introduction **1**
- 1.1 Research Problem . . . 2
- 1.2 Research Questions . . . 3
- 1.3 Objectives . . . 3
  - 1.3.1 General Objective . . . 4
  - 1.3.2 Specific Objectives . . . 4
- 1.4 Methodology . . . 4
- 1.5 Thesis Outline . . . 4
- 1.6 Contributions . . . 6

2 Related Work **7**
- 2.1 Social Media Messages in Mass Emergency . . . 7
  - 2.1.1 Classification, Extraction and Summarization during Emergency Situations . . . 8
  - 2.1.2 Crisis-Related Social Media Monitoring . . . 12
- 2.2 Event Detection Based on Locations . . . 14
- 2.3 Location Extraction in Social Media . . . 14
- 2.4 Summary . . . 16

3 Theoretical Framework **17**
- 3.1 Classification . . . 17
  - 3.1.1 Support Vector Machine Classifier . . . 18
  - 3.1.2 Decision Tree . . . 19
  - 3.1.3 Random Forest . . . 19
- 3.2 Clustering . . . 20
  - 3.2.1 Partitional Models . . . 21
  - 3.2.2 Density Models . . . 22
- 3.3 Normal Distribution . . . 23
- 3.4 Evaluation Methods and Metrics . . . 23
  - 3.4.1 Confusion Matrix . . . 24
  - 3.4.2 Clustering Evaluation . . . 25

4 Methodology **27**
- 4.1 Data Pre-Processing . . . 27
- 4.2 Signal Creation . . . 29
  - 4.2.1 Geographical Hierarchy . . . 29

4.2.2 Location Extraction . . . 30
4.3 Time-Window . . . 31
4.3.1 Features Extraction . . . 31
4.3.2 Determining Optimal Window Size . . . 32
4.3.3 Features Normalization . . . 33
4.4 Geographic Spread . . . 33

5 Experimental Setup **35**
5.1 Dataset Description . . . 35
5.2 Ground Truth . . . 36
5.3 Feature Selection . . . 37
5.3.1 Remove Redundant Features . . . 37
5.3.2 System's Historical Data . . . 38
5.4 Labeled Emergency Situation Events . . . 39
5.4.1 Under-Sampling . . . 39

6 Supervised Experimental Analysis **43**
6.1 Choosing a Machine Learning Classifier . . . 43
6.1.1 Support Vector Machine . . . 44
6.1.2 Random Forest . . . 45
6.2 Summary of the Supervised Results . . . 46

7 Evaluation **49**
7.1 Ground Truth . . . 49
7.1.1 Independent Analysis of Hierarchies . . . 49
7.1.2 Dependent Analysis of Hierarchies . . . 50
7.1.3 Geographic Spread Analysis . . . 51
7.2 On-line Evaluation . . . 54

8 Discussion and Conclusion **59**
8.1 Unsupervised Experimental Analysis . . . 61
8.1.1 Clustering Centroid Models . . . 61
8.1.2 Clustering Density Models . . . 62
8.1.3 Summary of the Unsupervised Results . . . 62
8.2 Final Comments and Future Work . . . 63

Bibliography **64**

Appendices **70**

A Tweet Object **71**

B Text pre-processing **73**

C Features Extraction **74**

D Printing Details to Remove Redundant Features **78**

E Results of the Supervised Approach **79**

F Results of the Unsupervised Approach **83**