

Contents

1. Introduction	1
1.1. Research Question and Challenges	3
1.2. Objectives	4
1.2.1. Main objective	4
1.2.2. Specific objectives	4
1.3. Contributions	4
1.4. Methodology	6
1.5. Outline of this work	6
2. Background and Related Work	7
2.1. Data Analysis Techniques	7
2.1.1. Classification	8
2.1.2. Clustering	12
2.1.3. Evaluation metrics	14
2.2. Event Identification	15
2.3. Prediction of Event Popularity	17
3. Modeling News Events and their Impact in Social Media	20
3.1. Pipeline to model Events	21
3.1.1. Retrieving Newsworthy Documents	22
3.1.2. Grouping similar news	24
3.2. Feature Extraction	25
3.3. Defining the Impact of an Event	25
4. Experimental Methodology	30
4.1. Building the Dataset	31
4.1.1. Collecting related posts for news on Twitter	31
4.1.2. Identifying Events	35
4.2. Data Cleaning and Validation	36
4.2.1. Detecting stopwords	37
4.2.2. Validation of Event Modeling	39
4.2.3. Events duration	40

4.3. Finding High Impact Events	42
4.4. Classification and Prediction	44
5. Characterization of News Events	47
5.1. Exploratory Analysis	47
5.2. Characterization of News by Impact	50
5.2.1. Information Forwarding characteristics	51
5.2.2. Interaction characteristics	54
5.2.3. Topical Focus characteristics	56
5.2.4. Attention characteristics	56
5.3. Classification and Prediction Results	59
6. Conclusions	62
6.1. Threats to Validity	63
6.2. Discussion and Future Work	63
6.2.1. Evolution of Events in Time	64
6.2.2. Geographic Characterization	64
6.2.3. Topical Features	65
6.2.4. Evaluation and Validation of Events	66
Bibliography	67
Appendices	71
A. News Sources	72
B. Plots for all Features	74
C. Results from Latent Dirichlet Allocation experiments	92