Título: EXPLORING THE TOPICAL STRUCTURE OF SHORT TEXT THROUGH PROBABILITY MODELS: FROM TASKS TO FUNDAMENTALS

Nombre: CAPDEVILA PUJOL, JOAN

Universidad: Universidad Politécnica de Catalunya

Departamento: Arquitectura de computadors

Fecha de lectura: 26/09/2019

Programa de doctorado: Programa de Doctorado en Arquitectura de Computadores por la Universidad Politécnica de Catalunya

Dirección:

> **Director:** JORDI TORRES VIÑALS

> **Codirector:** JESÚS CERQUIDES BUENO

Tribunal:

> **presidente:** JERÓNIMO HERNANDEZ GONZALEZ

> **secretario:** DAVID CARRERA PÉREZ

> **vocal:** MARK JAMES CARMAN

Descriptores:

> INTELIGENCIA ARTIFICIAL

> ANALISIS DE DATOS

> TECNICAS DE INFERENCIA ESTADISTICA

El fichero de tesis no ha sido incorporado al sistema.

Resumen: Recent technological advances have radically changed the way we communicate. Today¿s communication has become ubiquitous and it has fostered the need for information that is easier to create, spread and consume. As a consequence, we have experienced the shortening of text messages in mediums ranging from electronic mailing, instant messaging to microblogging. Moreover, the ubiquity and fast-paced nature of these mediums have promoted their use for unthinkable tasks. For instance, reporting real-world events was classically carried out by news reporters, but, nowadays, most interesting events are first disclosed on social networks like Twitter by eyewitness through short text messages. As a result, the exploitation of the thematic content in short text has captured the interest of both research and industry.
Topic models are a type of probability models that have traditionally been used to explore this thematic content, a.k.a. topics, in regular text. Most popular topic models fall into the sub-class of LVMs (Latent Variable Models), which include several latent variables at the corpus, document and word levels to summarise the topics at each level. However, classical LVM-based topic models struggle to learn semantically meaningful topics in short text because the lack of co-occurring words within a document hampers the estimation of the local latent variables at the document level. To overcome this limitation, pooling and hierarchical Bayesian strategies that leverage on contextual information have been essential to improve the

quality of topics in short text.
In this thesis, we study the problem of learning semantically meaningful and predictive representations of text in two distinct phases:
¿ In the first phase, Part I, we investigate the use of LVM-based topic models for the specific task of event detection in Twitter. In this situation, the use of contextual information to pool tweets together comes naturally. Thus, we first extend an existing clustering algorithm for event detection to use the topics learned from pooled tweets. Then, we propose a probability model that integrates topic modelling and clustering to enable the flow of information between both components.
¿ In the second phase, Part II and Part III, we challenge the use of local latent variables in LVMs, specially when the context of short messages is not available. First of all, we study the evaluation of the generalization capabilities of LVMs like PFA (Poisson Factor Analysis) and propose unbiased estimation methods to approximate it. With the most accurate method, we compare the generalization of chordal models without latent variables to that of PFA topic models in short and regular text collections.
In summary, we demonstrate that by integrating clustering and topic modelling, the performance of event detection techniques in Twitter is improved due to the interaction between both components. Moreover, we develop several unbiased likelihood estimation methods for assessing the generalization of PFA and we empirically validate their accuracy in different document collections. Finally, we show that we can learn chordal models without latent variables in text through Chordalysis, and that they can be a competitive alternative to classical topic models, specially in short text.