

Why Text Categorization?

- To **enrich text representation** (more understanding of text)
 - Text can now be represented in multiple levels (keywords + categories)
 - Semantic categories assigned can be directly or indirectly useful for an application
 - Semantic categories facilitate aggregation of text content (e.g., aggregating all positive/negative opinions about a product)
- To **infer properties of entities** associated with text data (discovery of **knowledge about the world**)
 - As long as an entity can be associated with text data, we can always use the text data to help categorize the associated entities
 - E.g., discovery of non-native speakers of a language; prediction of party affiliation based on a political speech