Variety

Classification and Clustering

Variety

- The Cluster System(UN, Red Cross…), - for humanitarian response for
  - Health, protection, food security…,
- *“This chapter describes methods for automatic text categorization, which allow us to make sense of heterogeneous, varied massages by sorting them into categories.”*
- Emergencies are rarely exactly the same, although there are commonalities.

Content Categories

Which categories should be used?

- Information needs,
  - What information does the user seek? It depends on the user,(humanitarian relief workers, policy makers,...)
- Capabilities of the system,
  - Certain tasks are hard for a computer to do. E.g: sarcasm, funny, literal or metaphorical.
- Capabilities of humans
  - Are bad at large numbers of categories.
  - Non-experts might have a even harder time separating on nuance.
- Availability of information,
  - All categories are not necessarily represented equally by the public.
  - Abundans of representation → fine grained typology.
  - Lesser quantities → coarse-grained typology.

Content Categories

eg: Existing Typologies by:

- Factual, subjective or emotional
- Information provided
- Information source
- Credibility
- Time
- Location

Information progression pattern in disasters(Olteanu et al. 2015)

- Caution and advice,
- Sympathy and support,
- Infrastructure damage and affected individuals
- Useful messages covering various topics
- Donations and volunteering

Supervised Classification

- Binary
  - Disjointed classes, can only be one of 2.
  - The message is either related to the event or not.

- Multiclass
  - Disjointed, can only be one of many.
  - The message during a tornado, could be one of: advice, donation, weather reports.

- Multilabel(tagging)
  - Can be many classes.
  - The message could be about about donation of food and cloth, rather then one or the other.

Supervised Classification

Elements of supervised classification systems:

- Training examples
  - A set of messages with known classes.
- Feature Selection
  - Methods used to pinpoint the aspects of the messages that best describes its desired classes.
- Learning algorithm
  - A model which tries to establish mappings that associates its inputs(a message) with the correct output(its class).
    - 1. Learning/training phase,
      - tunes internal parameters to best mimic the Training examples.
    - 2. Testing/labeling phase
      - Tries to map inputs with outputs, based on the current internal parameters.
    - SVM, Trees….
- Evaluation Metrics
  - Test the trained machine learning model how effective it is at mapping inputs to outputs(finding the correct classes for the associated messanges).
    - Precision and Recall, ROC,...

Unsupervised Classification / Clustering

“Exploratory methods that search for pattern or structure unlabeled data.”

- Hard clustering
  - A item(message), can only be part of one cluster of items.
  - k-means

- Soft clustering
  - A item(message), can be apart of many clusters of items.
  - LDA

- Clustering Granularity
  - Coarse-granularity
    - Tries to separate
    - Few large clusters
  - Fine-granularity
    - Many small clusters.
    - Clusters together messages that “basically” means the same thing.

Research Problems

- Adapting classification models to new situations
  - Being able to reuse old well established classification models, but with alterations with in the current topic.
- Interactive taxonomy design
  - Topics are often hard to establish prior to a emergency. Therefore an adaptive design that expert could alter on the go to better fit the situation is in need.
- Ranking
  - Rank importance of messanges.