University of BRISTOL

1.1 Introduction to Text Analytics

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

What is ‘Text Analytics’?

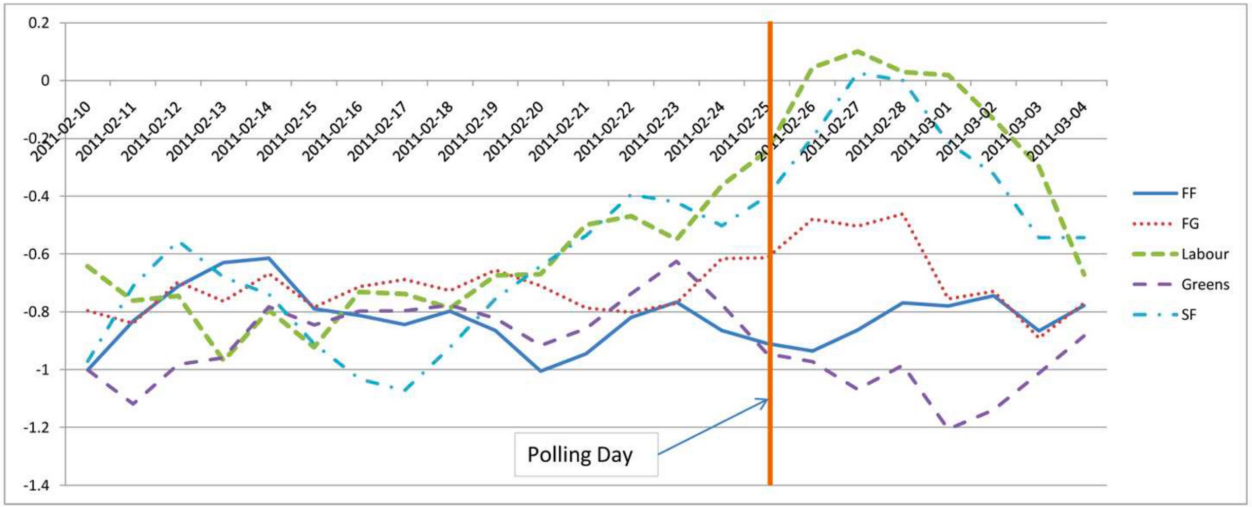
- Text as data
- Transform **unstructured** text to **structured** data
- Uncover facts, events, opinions, and other information from text
- Combine numerous sources to derive new insights
- Analyse and visualise topics and trends

Why is Text an Important Data Source?

- Huge amounts of information on the web, such as:
  - Wikipedia – an expansive source of information
  - Social media – opinions, eyewitness acounts, …
- Free text records in databases often contain valuable insights, e.g., why a customer cancelled a contract.
- Within an organisation, reports and case logs explain what happened and why
- Language can express all manner of relationships and reasoning that are difficult to uncover from structured data

Example Tasks: Sentiment Analysis

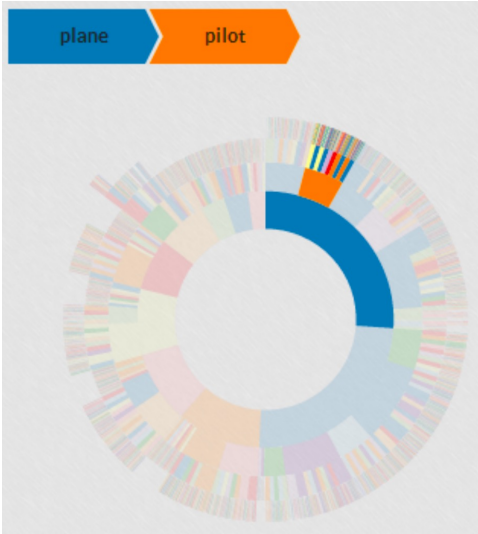
- Identify positive or negative attitudes or opinions
- **Classify** a sentence or document to a category label, “+ve” or “-ve”
- Example uses: understanding political opinions over time



From: Bermingham et al. (2011)

Example Tasks: Topic Modelling

- Identify common topics across multiple documents
- Example uses:
  - Finding sets of documents with common themes
  - Summarising the topics in a set of documents
  - Tracking topic trends over time



Smith et al. (2014).

Example Tasks: Information Extraction

- Extract structured information from text, such as:
  - Entities = people, places, organisations, specific events…
  - Events
  - Relations between entities and events
- Example uses: populating a knowledge base with facts

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

FARE-RAISE ATTEMPT:		
	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

How Can We Solve These Tasks?

- Natural language processing (NLP) – computational methods that are the core of text analytics
- Linguistics – understand how language expresses meaning
- Machine learning – methods for learning from data

Chapter 1 of Introduction to Natural Language Processing, Eisenstein, J. (2019).

What Makes Text Special?

- Text data is **discrete.** Discrete units are combined in sequences to form meaning.
- Many observations are **rare**, many possible sentences are never observed in any given dataset.
- Text is **compositional:** words combine into phrases, which combine to form sentences, and so on.
- Ambiguity, errors and variations in the way people use language also present major challenges.

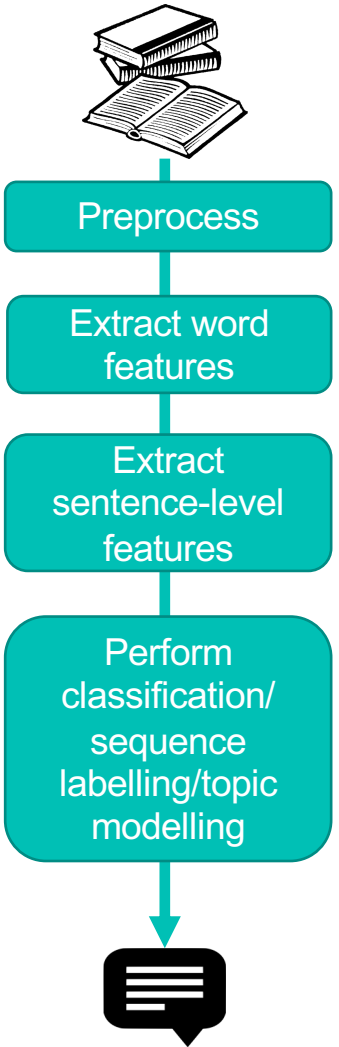
Ways to View Meaning in Language

Umashanthi interviewed Ana. She works for the college newspaper.

- Relational
  - Relationships between words represent meaning
  - E.g., synonyms, categories, …
- Compositional
  - The meaning of larger units is formed by combining smaller units
  - E.g., sentences from phrases, words from suffixes, prefixes and stems
- Contextual (or ‘distributional’)
  - We can understand a word from its context
  - The context of a word alters its meaning

NLP Pipelines

- NLP pipelines apply a series of processing steps to an input text
- Each step takes the outputs of earlier steps as its inputs
- Early steps extract low-level features, e.g., tokens
- Further steps may analyse the structure of the text
- The final stage uses these features to compute the system's outputs



Ethical Considerations

- Privacy and freedom of speech: whose data are we processing, and does doing so restrict their privacy or freedom?
- Labour: who created the data that we're using and do they benefit from our technology?
- Bias: technology can amplify societal biases, so have we done enough to identify and address potential problems? Do we serve different communities (e.g., speakers of different languages) equally?

Summary

- Text analytics extracts insights and structured data from text data using NLP.
- Example tasks: sentiment analysis, topic modelling and information extraction.
- Meaning: relational, compositional and contextual views.
- NLP pipelines extract features sequentially to perform complex tasks.
- Developing text analytics systems requires ethical considerations regarding privacy, sources of labour and bias.

Reading

- Jacob Eisenstein, *Introduction to natural language processing*, 2019.
  - Available in the library
  - Or see the free online draft
- Dan Jurafsky, James H. Martin, *Speech and Language Processing (3rd edition draft):*
  - *https://web.stanford.edu/~jurafsky/slp3/*
  - 2nd edition is available in the library
- **Reading for this video: Chapter 1 of Eisenstein**