

Tips for Final Project

Jaime Arguello

INLS 613: Text Data Mining

jarguell@email.unc.edu

Examples

- Given a Yelp review, predict its star-rating (1-5)
- Given a wine review, predict its star-rating (1-10)
- Give a drug review, predict its sentiment
- Given a research article, predict whether it belongs in a specific “systematic review”
- Given tweets with ticker references, predict whether the S&P500 will go up/down in time $t+1$
- Given a tweet, predict whether the author believes in climate change (yes/no)
-

Tips (1)

- Ask interesting research questions that are grounded on:
 - ▶ Theory
 - ▶ Intuition
 - ▶ Error analysis using a simple baseline

Example from Sentiment Analysis of Movie Review

- **Baseline:** simple unigram representation
- What are language phenomena not captured by a simple unigram representation?
- **Rephrase:** What are contextual factors that influence whether a word predicts positive, negative, or neither?

Example from Sentiment Analysis of Movie Review

- **Rephrase:** What are contextual factors that influence whether a word predicts positive, negative, or neither?
 - ▶ POS (“love_NOUN” vs. “love_VERB”)
 - ▶ Negation (“not”)
 - ▶ Modal verbs (“could”, “should”, “would”)
 - ▶ Plot description vs. opinion
 - ▶ Movie genre (“comical”, “ridiculous”)
 - ▶ MPAA rating (“violence”)

Examples

- Given a Yelp review, predict its star-rating (1-5)
- Given a wine review, predict its star-rating (1-10)
- Give a drug review, predict its sentiment
- Given a research article, predict whether it belongs in a specific “systematic review”
- Given tweets with ticker references, predict whether the S&P500 will go up/down in time $t+1$
- Given a tweet, predict whether the author believes in climate change (yes/no)
-

Tips (2)

- Beware of uninteresting confounds
 - ▶ # of positive tweets predicts strong consumer confidence
 - ▶ % of positive tweets predicts strong consumer confidence

Tips (3)

- Evaluate using a sensible metric

Examples

- Given a Yelp review, predict its star-rating (1-5)
- Given a wine review, predict its star-rating (1-10)
- Give a drug review, predict its sentiment
- Given a research article, predict whether it belongs in a specific “systematic review”
- Given tweets with ticker references, predict whether the S&P500 will go up/down in time $t+1$
- Given a tweet, predict whether the author believes in climate change (yes/no)
-

Tips (4)

- Compare your approach against all “degenerate” baselines

Tips (5)

- Your goal is not to solve the problem
- Your goal is to learn about the problem and provide directions for others to follow