# Text Mining in Social Media

## Deception Detection

Javier Fernández-Bravo Peñuela

Master in Big Data Analytics

Universitat Politècnica de València

# Main challenges

- Author profiling
- Deception detection
  - Irony/humor detection
- Arabic Language Processing

# Related areas

- Text classification
- Natural Language Processing

# Phases and methods

NLP techniques:

- Stemming
- Lemmatization
- Tokenization
- Tagging

ML classification:

- Decision trees
- Naive Bayes
- Support Vector Machines
- Hidden Markov Models

# Techniques to be applied

- Normalization (capitalization, numbers, and punctuation)
- Tokenization
- Removal of stopwords (both English and Arabic)
- Definition of dataset vocabulary (most 1000 frequent words)
- Bag of Words

# Special proposal

- Different classifiers for news and tweets datasets
- Extraction of additional features related to the text structure (ex., length)
- Extraction of additional features related to Twitter texts: hashtags, user mentions, emojis

# Text classification

- Support Vector Machines
  - Linear kernel
  - Test across multiple combinations of hyperparameters in order to optimize the classifier’s model tuning

# Results

- K-fold cross validation (3 folds training, 1 fold evaluation)
- Performance of the best classifier on evaluation data fold, in absence of the test dataset

| Model | News | Twitter |
|---|---|---|
| Baseline | 0,58 | 0,61 |
| Intermediate model | 0,70 | 0,64 |
| Final model with ad-hoc Twitter features | 0,70 | 0,77 |