(Spam *vs.*) Forty Years of Machine Learning for Text Classification

David D. Lewis, Ph.D.

Independent Consultant

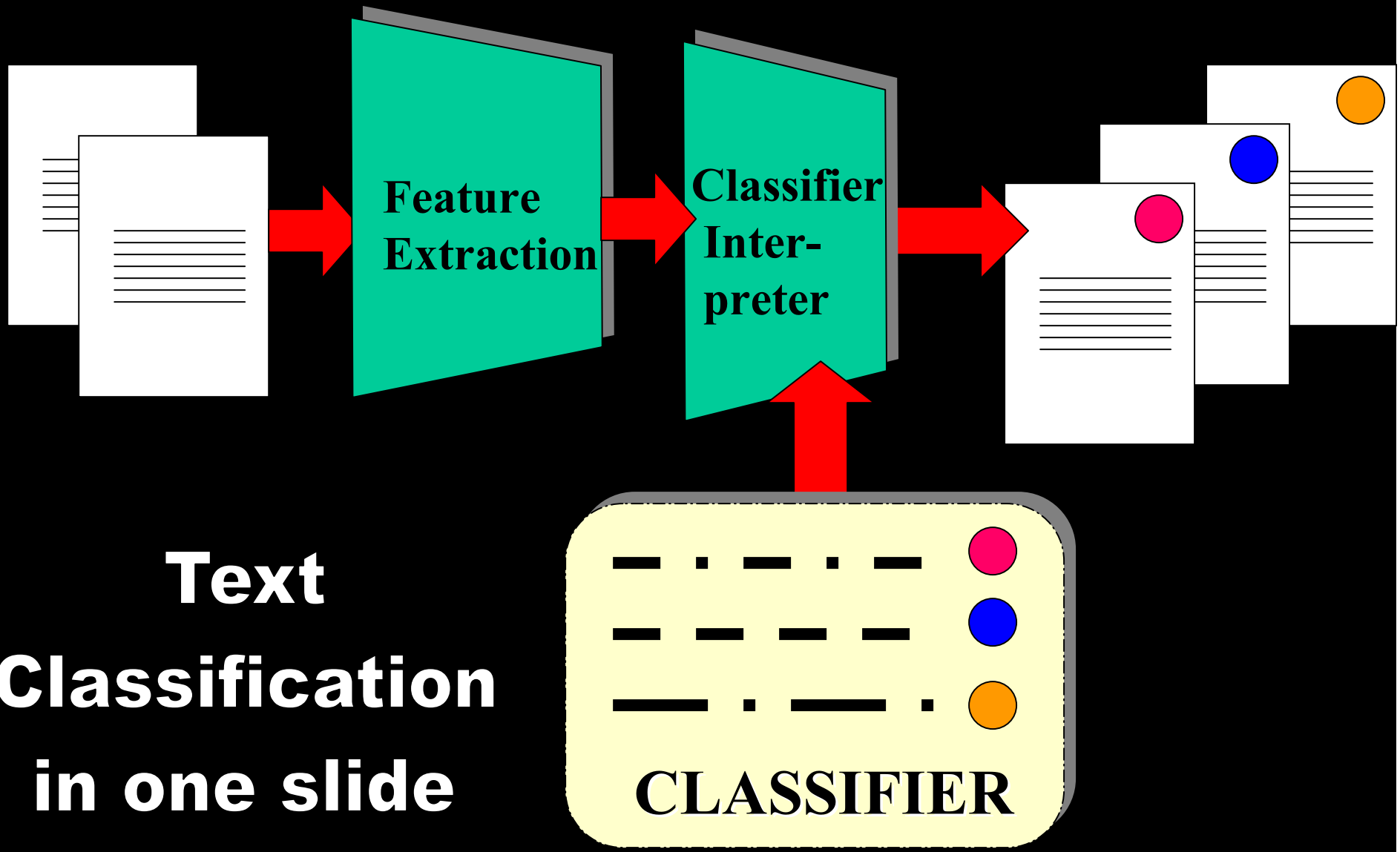
Chicago, IL, USA

Dave@DavidDLewis.com

www.DavidDLewis.com

Presented at the 1st Spam Conference, Cambridge, MA, 17-Jan-03

Copyright (c) 2003 David D. Lewis



Text Classification in one slide

Supervised Learning of Text Classifiers

- A supervised learning program produces a classifier given input/output pairs:

  (Doc1, **X**), (Doc2, **Y**), (Doc3, **X**), (Doc4, **X**)…

  – Input: text represented as binary/numeric features

  – Output: class (e.g spam vs. not-spam (ham))

- Why?: algorithms better than humans at producing formulae for combining evidence

Copyright (c) 2003 David D. Lewis

Text Representation

- Human language -> feature vectors
- Term weighting, feature selection, collocations, multiple reps help
- Most variations among *task-independent* text processing have little impact:
  - Tokenization, stemming, NLP,...
  - Clustering, LSI, ICA, Kohenen nets,...
  - Wordnet, Roget's Thesaurus,...

Copyright (c) 2003 David D. Lewis

Character N-Grams Phrases?

- If all text representations about the same, maybe pick the most robust:
  - Downcase
  - Remove markup (“eye space”), punctuation, numbers, spaces(!)
  - Frequent, sparse character n-gram phrases?
- Defeats “M A K E * M-O-N-E-Y”, etc.
- Intended to be only one of several reps

Copyright (c) 2003 David D. Lewis

Feature Engineering

- *Task-specific* feature engineering helps a lot
  - Construction of text & nontext features
  - Ubiquitous in operational TC
- Good features are a learner's best friend
- Cautions:
  - Don't tune features on same data used for learning
  - Avoid learners that lock on one good feature

Copyright (c) 2003 David D. Lewis

Future of Text Representation?

- Current filters pay somewhat more attention to structural than linguistic features
  - Forged headers, broken markup, etc.
- But pressure of filtering will (slowly!) force spam to look more legit
- Text content will be key as this happens:
  - MLM is MLM is MLM

Copyright (c) 2003 David D. Lewis

Classifier Form

- What math function combines evidence?
- Weighted better than Boolean (“rules”)
  - Handles graded, probabilistic classes
  - Moderate advantage on effectiveness
  - Large advantage on robustness?
- Linear as effective as nonlinear
  - And has fast, simple learning algorithms
  - Better to create nonlinear terms by hand

Copyright (c) 2003 David D. Lewis

Learning Algorithm

- (Over?)emphasized in TC research
  - Naive Bayes one of hundreds of learning algorithms just for linear models
  - Effectiveness only one of many criteria
    - Uneven costs widely dealt with
- More important is *understanding* the algorithm you use
  - e.g. How interacts with engineered features

Copyright (c) 2003 David D. Lewis

Training Data Basics

- More data better than less
- Manually labeled (much) more useful than unlabeled
  - Despite progress on using unlabeled data
- Accurately labeled
- Broad coverage of range of inputs
- Same features as classifier will encounter

Copyright (c) 2003 David D. Lewis

Selecting Training Data

- Data collection has emphasized spam
- Privacy means little ham available
  - Exchange summary stats (only Naive Bayes)
  - Anonymized examples (limits features)
  - Tune thresholds by hand
- Likely to fail as spam tries to look legit
  - Will need full linguistic content of ham

Copyright (c) 2003 David D. Lewis

Active Learning

- Choose training examples at current classifier boundary
  - Mistakes, near misses
- One actively selected example worth 100's of random ones
- Iterative approach particularly powerful:
  - Train, select, label, repeat
- (Related methods to build evaluation sets)

Copyright (c) 2003 David D. Lewis

Active Learning & Privacy

- I have 100,000+ saved emails
  - Won’t share with strangers
  - Won’t manually classify (well, actually I did...)
- Send me a program:
  - I run it on my mail archive
  - It identifies 10 boundary cases
  - I decide which I’m willing to share
- Repeated over many volunteers would be almost as effective as everyone sharing all data!

Copyright (c) 2003 David D. Lewis

Summary

K-2

Kindergarten, 1st, and 2nd Grades.
Click on the bear to enter the world of fun and learning!
Math Language Arts

3-4

Grades 3 and 4.
Learning IS fun!
Click on the butterfly to enter.
Math Language Arts

Copyright (c) 2003 David D. Lewis

Evaluation

- Much more careful data collection needed for evaluation than training
- Goodman talk made many very good points

Copyright (c) 2003 David D. Lewis

Advertisements

- Operational Text Classification workshop:
  - http://www.DavidDLewis.com/events/
- Contact me (Dave@DavidDLewis.com) re:
  - Low volume discussion list for researchers/practitioners in text classification
  - Planned edited collection on practical experiences with text classification
  - Consulting

Copyright (c) 2003 David D. Lewis