oiioiioii oxford
oiioiioii internet institute
oiioiioii university of oxford

ÅBO AKADEMI
Information Studies

Sentiment analysis tasks and methods

Mike Thelwall
University of Wolverhampton, UK

Statistical Cybermetrics
Research Group

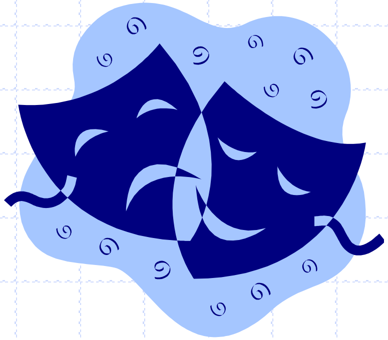
CYBEREMOTIONS

UNIVERSITY OF
WOLVERHAMPTON

Contents

- Types of sentiment analysis task
- Standard machine learning methods
- Linguistic algorithms

Terminology and problems



- Sentiment Analysis (SA), AKA Opinion Mining, is the task of automatically detecting sentiment in text
  - Active research area since ~2002
  - Standard part of online market research toolkits
  - Commonly used for automatic processing of large numbers of texts to identify opinions about products or brands
- *Opinions* are personal judgements about something
  - It is good. It is bad. It is expensive.
  - *Subjective* text contains opinions; *Objective* text states only facts.
- *Sentiments* are expressions of emotion or attitude or opinion
  - It is good. It is bad. It is expensive. I like it. I am happy. I am depressed. I am angry at you.
- Sentiment analysis is sometimes thought of as the prediction of people's private/internal states from text

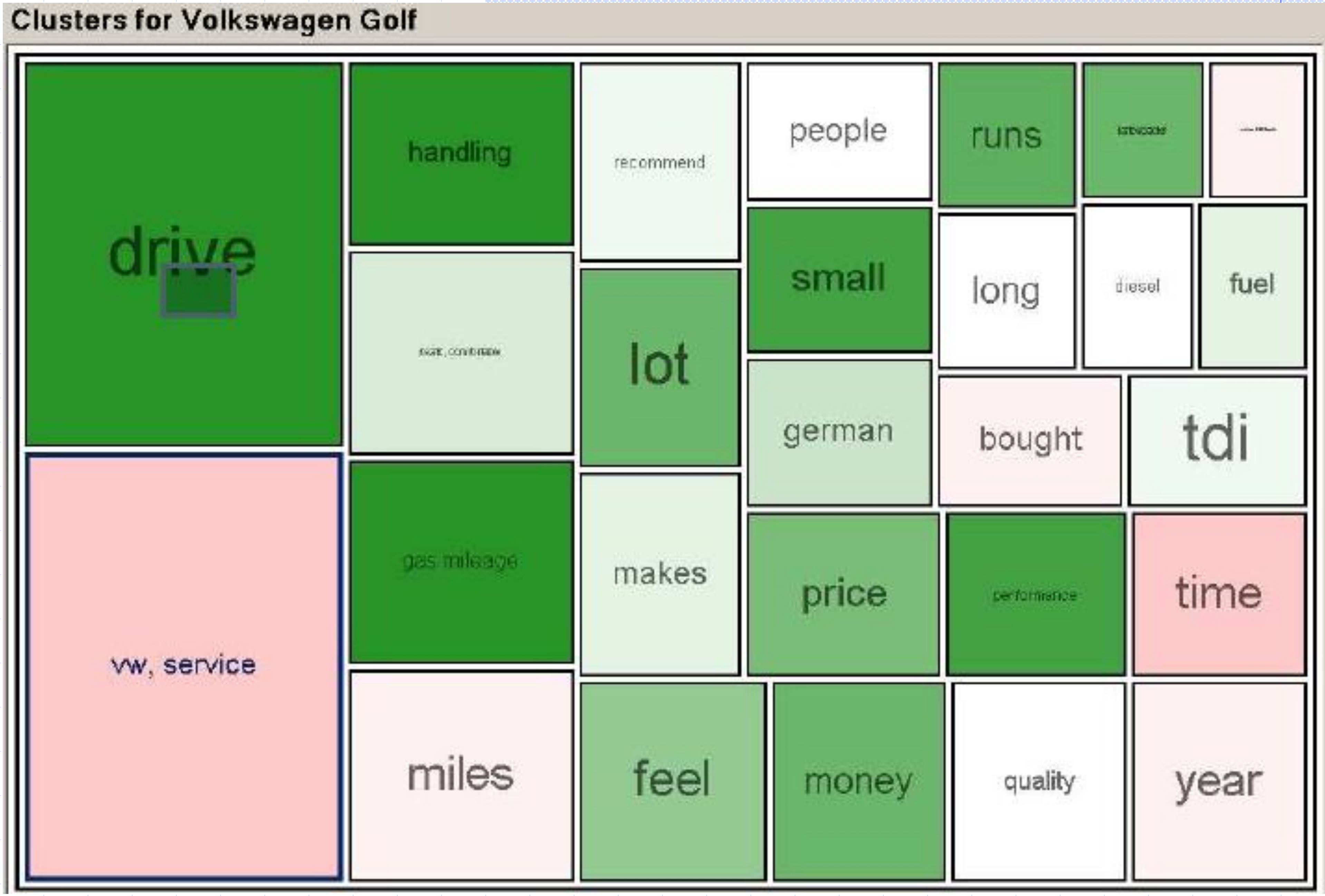
http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html

Opinion Mining Applications

- Identify unpopular features of BMWs
  - Automatic analysis of thousands of comments in BMW car forum
  - Identify features *and* sentiment
- Identify if a new computer is popular
  - Automatic analysis of all blogs
  - Compare to results for other computers
- Identify impact of TV advertising campaign
  - Automatic analysis of all blogs
  - Identify and detect sentiment in product mentions

Commercial sentiment analysis goals

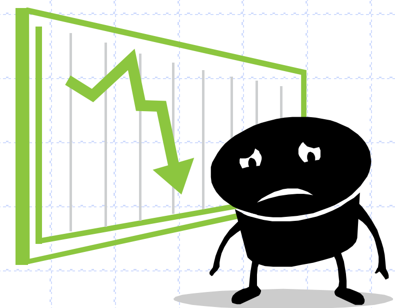
- Determine overall opinions about a ***product***
  - E.g., the M90 phone is excellent.
  - E.g., the M90 is expensive but excellent.
- Determine opinions about ***parts*** of a product
  - E.g., the screen of the M90 is too small but its weight is very light.
  - I love the steering wheel on the new Picasso!



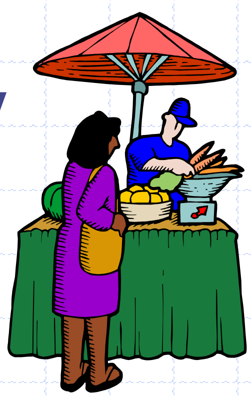
Gamon, et. al. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science, 3646*, 121-132.

Commercial sentiment analysis goals

- Determine *changes* in overall customer brand opinion (e.g., daily proportions of positive/negative comments)
  - In response to advertising
  - As routine monitoring

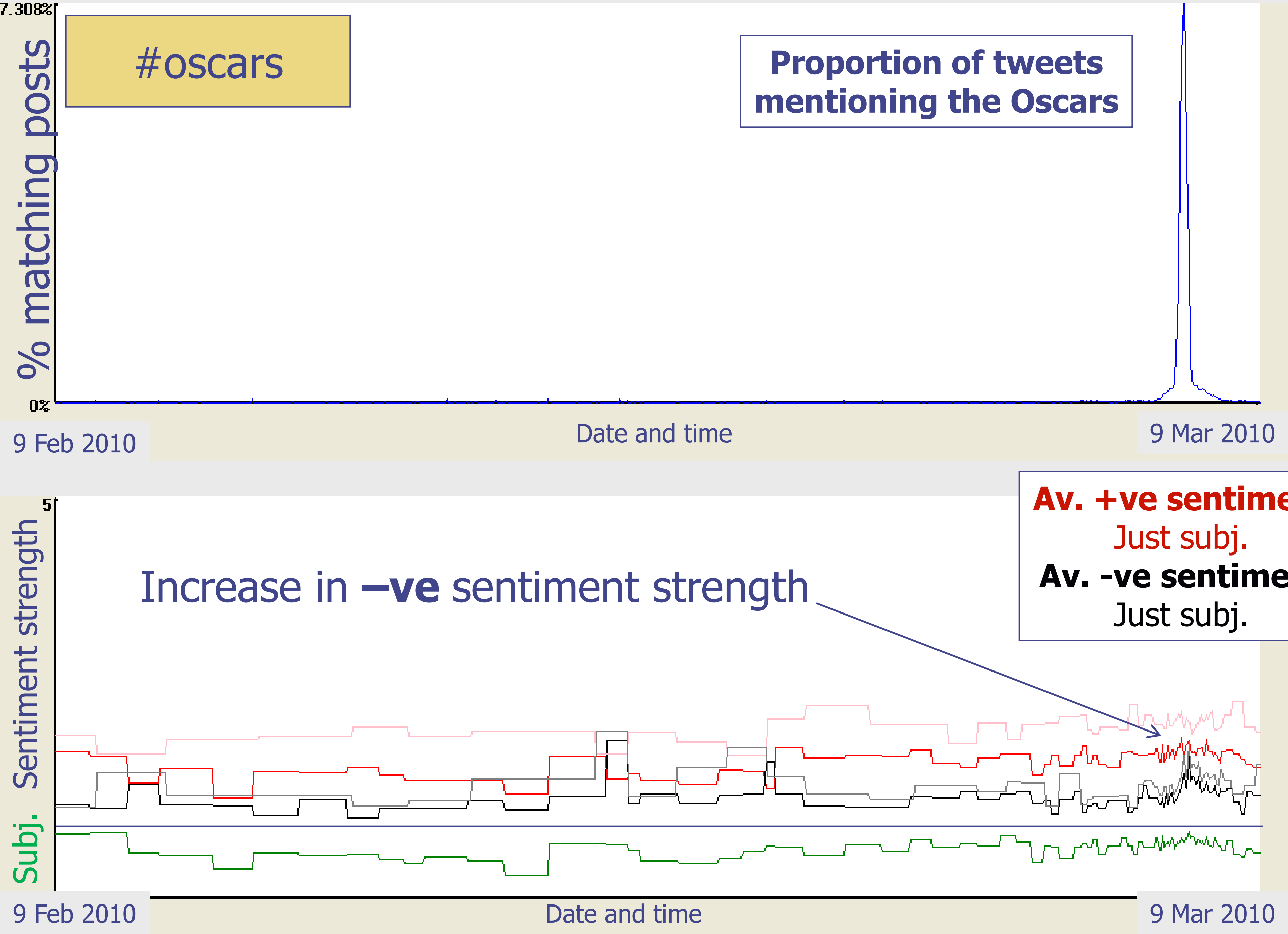


- Identify individual unhappy customers
  - E.g., identify Tweets that mention the brand and are negative
  - *Endnote web is driving me mad, argggggh!!!*



Social science sentiment analysis goals

- Track trends in sentiment over time (see next slide)
- Identify changes in sentiment
- Discover patterns in sentiment use in a communication medium
  - E.g., gender, age, nationality differences
  - Do women/Russians use more sentiment?



#oscars
Proportion of tweets mentioning the Oscars
7.308%
0%
% matching posts
9 Feb 2010
Date and time
9 Mar 2010
Av. +ve sentime
Just subj.
Av. -ve sentime
Just subj.
Increase in –ve sentiment strength
5
Sentiment strength
Subj.
9 Feb 2010
Date and time
9 Mar 2010

Types of sentiment analysis task 1: Subjectivity detection

- Detecting whether a text is opinionated/ subjective or neutral/ objective
- Binary decision
- Can use machine learning
- Does not classify polarity

```
This phone is very cheap.
This phone costs 200 roubles.
I love the phone.
```

Types of sentiment analysis task 2: Polarity detection

- Detecting whether a subjective text is positive or negative
- Binary decision
- Can use machine learning

```
This phone is very cheap.
It is lovely.
I am frustrated with the phone.
```

Types of sentiment analysis task 3: Sentiment strength detection

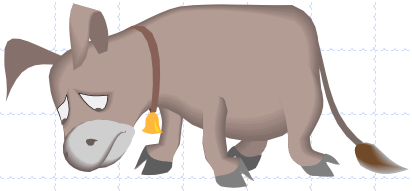
- Measuring the strength of sentiment in a text
- Scale ratings – many different ones used
- E.g.,
  - strong negative 1-2-3-4-5-6-7-8-9 strong positive OR
  - 1-2-3-4-5 negative & 1-2-3-4-5 positive

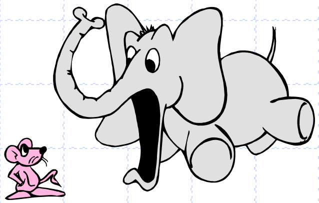
The car is very good.

I am tired but happy.

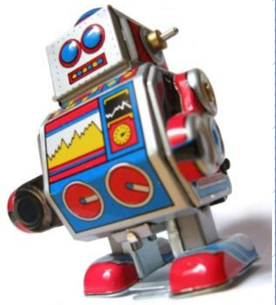
Types of sentiment analysis task 4: Multiple sentiment detection

- Detecting a range of emotions
  - E.g., happy, sad, angry, depressed, excited
- Is harder and some emotions are rare in text.

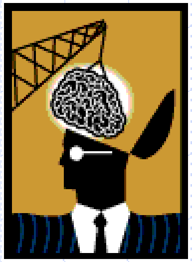




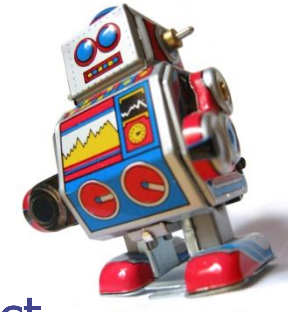
2. Machine learning



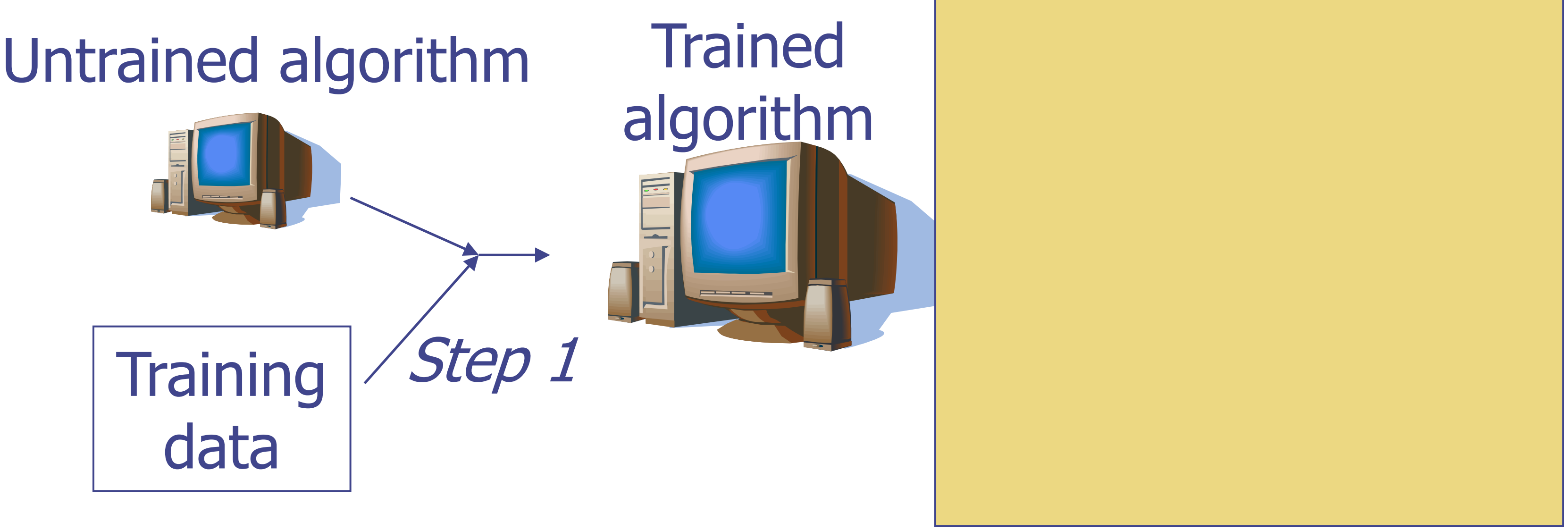
- Machine learning algorithms typically have a variety of parameters that can be "learned"
- Input set of human-classified texts
- Algorithm adjusts its parameters to perform well on the human-classified texts
  - Should also be accurate on *similar* new texts



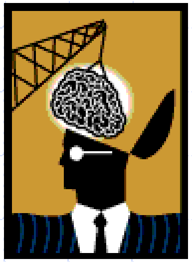
Machine learning overview



- Training data – (typically) human-annotated with the correct sentiment values and used for training the algorithm
- Test data – identical to the above except used for testing the trained algorithm to see how accurate it is



Training example

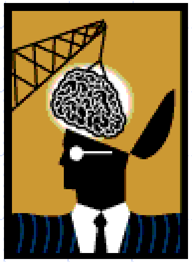


- Features : anna, hate, i, love, you
- d1 feature vector: (1,1,1,0,0)
- d2 feature vector: (1,0,0,1,1)
- An algorithm is told d1 is negative and d2 is positive: what will it learn?
-

d1 I hate Anna.

d2 I love you.

Training example 2



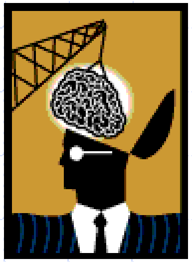
- What will the algorithm learn now?

d1 I hate Anna.

d2 I love you.

d3 I love Anna.

Types of machine learning algorithm

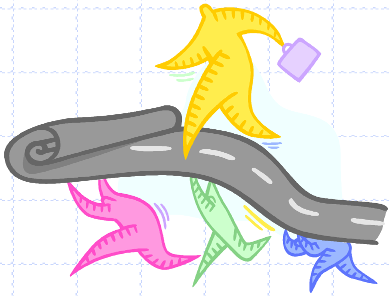


- Many generic and many highly tailored machine learning algorithms
- For text analysis there is an important distinction between types:
  - Linguistic – use grammatical and other knowledge about language to 'understand' the text analysed (e.g., SentiStrength)
  - Non-linguistic – use brute force methods that do not incorporate linguistic knowledge (e.g., with feature vector inputs)

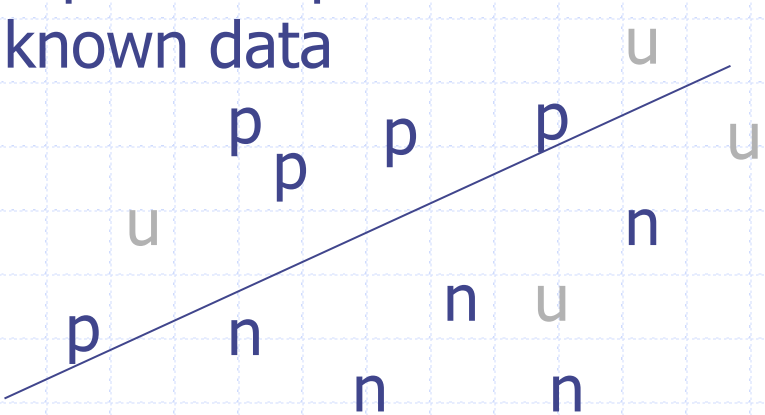
Non-linguistic algorithms

- General mathematical methods incorporating abstract intuitions about how to learn to guess correct sentiment value from training data
- The algorithm makes its judgement based solely on the feature vectors

Support Vector Machines



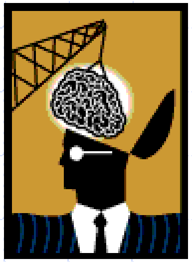
- Popular and powerful
  - maps the feature vectors into a high-dimensional space in a clever way
  - finds a hyperplane (a bit like a straight line) that separates the training data into two different classes as well as possible
  - uses the same hyperplane to predict the classes of the test data or unknown data



Support Vector Machines - Example

Find the hyperplane

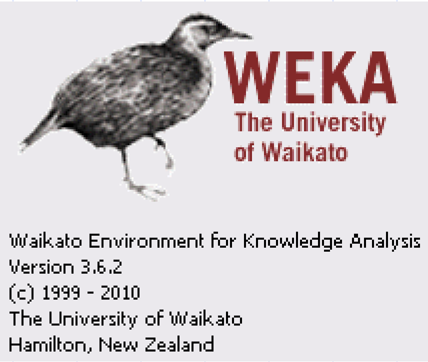
Other generic machine learning algorithms



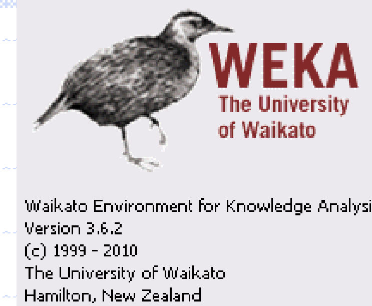
- Naïve Bayes – makes simple probability assumptions
- Rule generators – e.g. finds simple rules like "If document contains "love" and doesn't contain "hate" then classify it as positive"
- Genetic algorithms
- Logistic regression
- Decision tables
- Boosting algorithms
- Multilayer perceptron
- Many more, and each one has many variations and parameters ☹

Practical advice

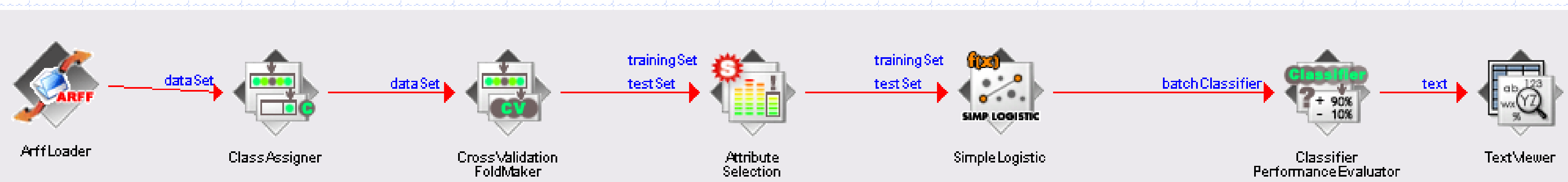
- Use Weka with many machine learning algorithms to run tests and develop a system (no programming needed) www.cs.waikato.ac.nz/ml/weka/
  - For text analysis, need to write code to convert data into feature vectors
- Or use text-specific analysis packages like GATE that focus more on natural language processing (gate.ac.uk)
- OR SVMLight (free online, fairly easy to use)



Weka

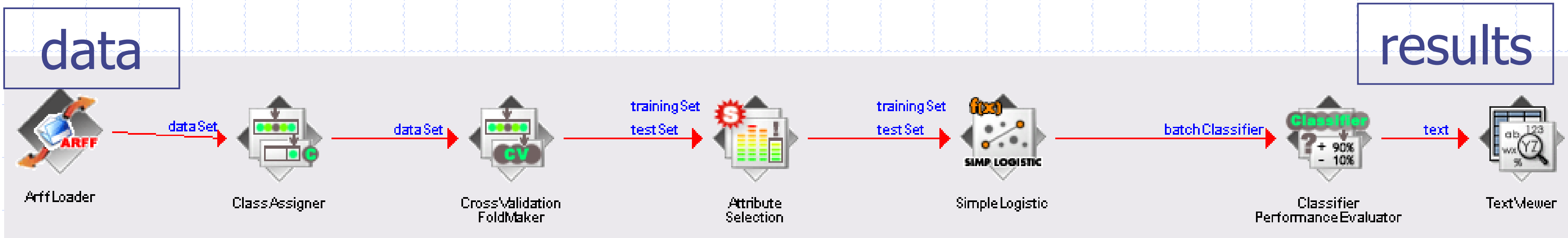


- Contains many components that can be built into processing pipelines
- Can use in five different ways
  - Explorer – load data and try different algorithms on it (not large data sets)
  - Experimenter – set up large-scale experiments with different algorithms and data
  - KnowlegeFlow – connect together multiple algorithms on the fly
  - Command line interface - one algorithm at a time
  - Java programs – API – for systematic and customised testing



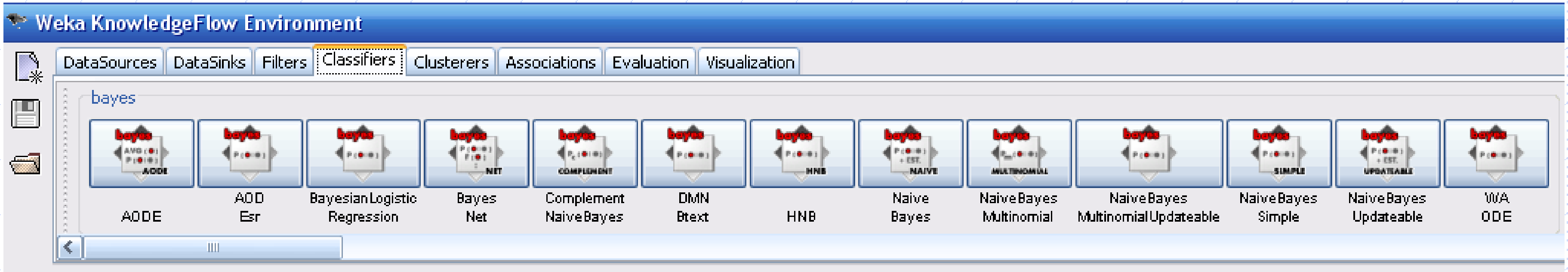
Sample Weka process

1. Load data file (add data file loading component to interface; enter name of data file)
2. Mark one of the data columns as "correct" or the class to be predicted
3. Split the data into training and testing sets
4. Train the ML algorithm to the training data, evaluate it on the testing data
5. Calculate accuracy statistics on the results



Weka 2

- Many different options
- Takes time to get used to
- Is very powerful and flexible
- Need ML understanding to use



Linguistic algorithms

- Incorporate additional grammatical and other information about language
- Typically use a scoring function to predict sentiment
- Tend to be more accurate but take much more time to run
- Examples of additional power:
  - The word 'like' can be positive (verb) or neutral (preposition) – linguistic techniques can disambiguate the two senses.
  - The words 'hate', and 'hated' have the same lexical root, and a similar meaning to 'loathe' and 'loathed'
  - 'not' often reverses the meaning of subsequent words
  - there are many idioms that have special meanings
  - sarcasm has special meanings
- Linguistic knowledge of the possible meanings of words can give algorithms a head start
  - E.g., SentiWordNet lists many classes of positive and negative words

Example - SentiStrength

- I love my Lada

->

- I do not hate traffic.

->

Linguistic resources

- Part of speech tagger
  - Predicts use of any given word
- Sentiment resource or lexicon
  - E.g., SentiWordNet = network of groups of sentiment words and meanings
- Chunker – identifies sentences and phrases
- Standard toolkits include Gate and LingPipe

SentiWordNet

- Example of a linguistic resource
- Based on WordNet
  - Database of word meanings and relationships
  - Can use to find words similar to any given word
  - Uses synsets – groups of semantically equivalent words
  - ~ 150,000 words in ~ 115,000 synsets
- SentiWordNet http://sentiwordnet.isti.cnr.it/
  - Assigns three probability-like scores to each WordNet Synset: for objectivity, positivity and negativity
  - The scores are based on an estimation algorithm
  - Powerful resource for estimating the sentiment of individual words
  - Needs linguistic processing of source text to match words to synsets.
    - E.g., to disambiguate the different 'like' synsets

Unsupervised algorithms

- An algorithm is supervised if it requires a training stage
- An algorithm is unsupervised if it requires no training
- An algorithm is semi-supervised if it has a limited training stage
- Machine learning algorithms tend to be supervised or semi-supervised
- Linguistic algorithms are often unsupervised = no need for training data

Summary

- There are several different sentiment analysis tasks and many different applications of sentiment analysis
- Machine learning is based upon algorithms that learn to solve classification or clustering problems from human-coded examples
- Linguistic algorithms use knowledge of language to improve performance, but may be less customisable to specific domains (see later)

Bibliography

- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 1*(1-2), 1-135.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
- Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science, 3646*, 121-132.