# Lecture 7: Sentiment Analysis

# Lexicon-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says “justice” vs “efficiency”)
- General dictionaries: WordNet, LIWC, MFD, etc.

# Measuring uncertainty in macroeconomy

> Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

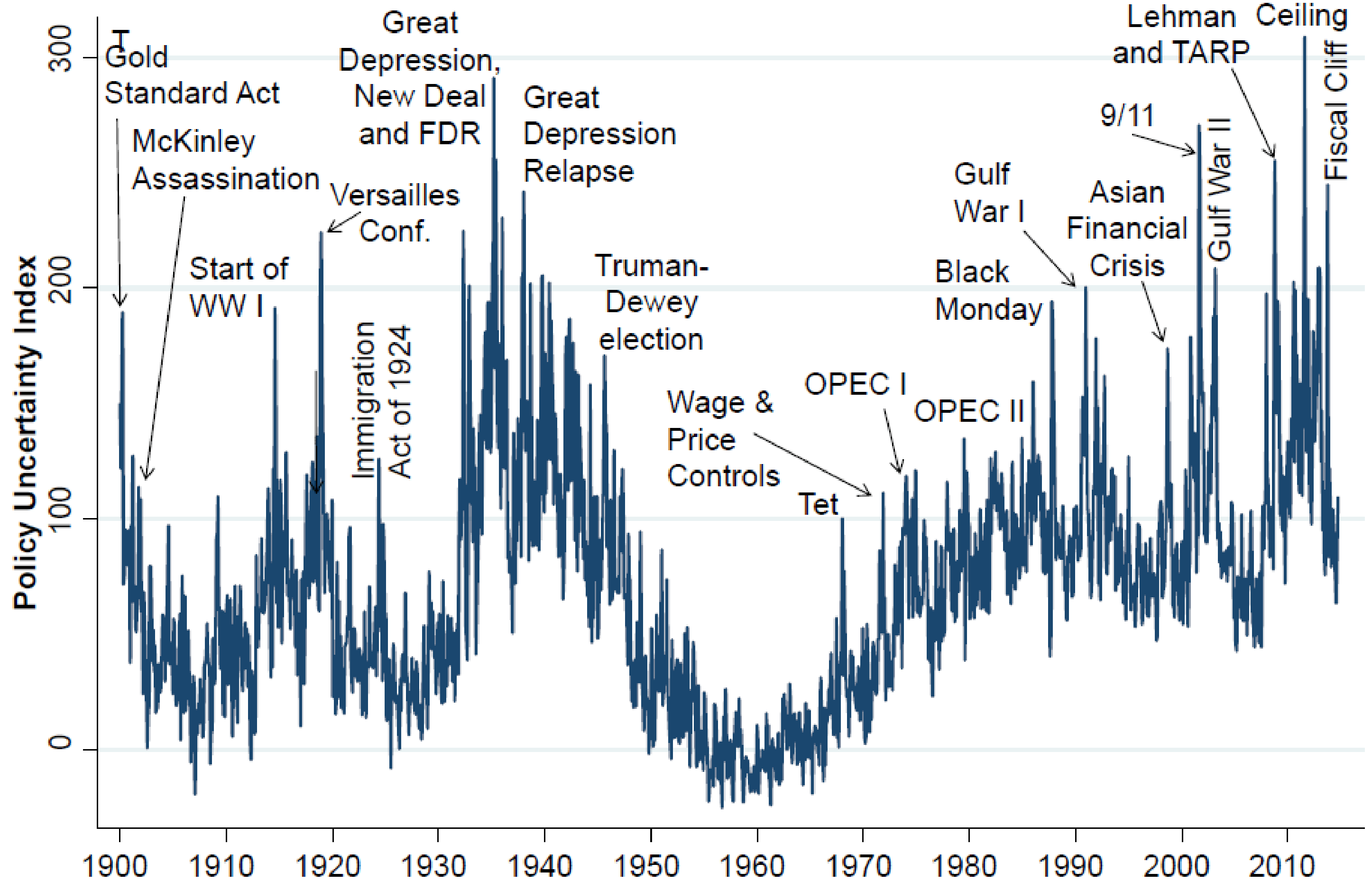
1. Article contains “uncertain” OR “uncertainty”, AND
2. Article contains “economic” OR “economy”, AND
3. Article contains “congress” OR “deficit” OR “federal reserve” OR “legislation” OR “regulation” OR “white house”

Normalize resulting article counts by total newspaper articles that month.

- but see Keith et al (2020), showing some problems with this measure (https://arxiv.org/abs/2010.04706).

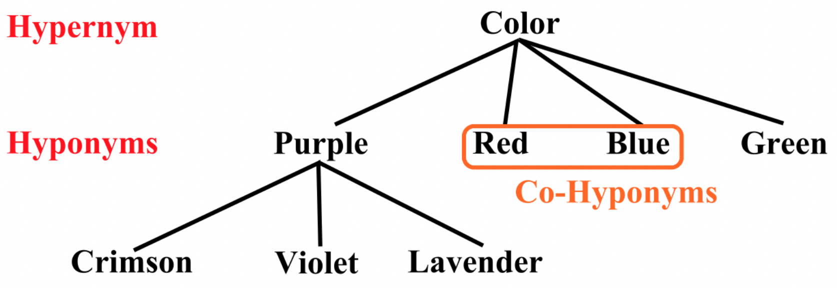
# Measuring uncertainty in macroeconomy

Figure 2: US Historical Index of Economic Policy Uncertainty



# WordNet

- English word database: 118K nouns, 12K verbs, 22K adjectives, 5K adverbs
- Synonym sets (synsets) are a group of near-synonyms, plus a gloss (definition).
    - also contains information on antonyms (opposites), holonyms/meronyms (part-whole).
- Nouns are organized in categorical hierarchy (hence "WordNet")
    - "hypernym" – the higher category that a word is a member of.
    - "hyponyms" – members of the category identified by a word.



# WordNet

The noun “bass” has 8 senses in WordNet.
1. bass$^1$ - (the lowest part of the musical range)
2. bass$^2$, bass part$^1$ - (the lowest part in polyphonic music)
3. bass$^3$, basso$^1$ - (an adult male singer with the lowest voice)
4. sea bass$^1$, bass$^4$ - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass$^1$, bass$^5$ - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass$^6$, bass voice$^1$, basso$^2$ - (the lowest adult male singing voice)
7. bass$^7$ - (the member with the lowest range of a family of musical instruments)
8. bass$^8$ - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

**Figure 19.1** A portion of the WordNet 3.0 entry for the noun *bass*.

# WordNet Supersenses (Word Categories)

| Category | Example | Category | Example | Category | Example |
|---|---|---|---|---|---|
| ACT | *service* | GROUP | *place* | PLANT | *tree* |
| ANIMAL | *dog* | LOCATION | *area* | POSSESSION | *price* |
| ARTIFACT | *car* | MOTIVE | *reason* | PROCESS | *process* |
| ATTRIBUTE | *quality* | NATURAL EVENT | *experience* | QUANTITY | *amount* |
| BODY | *hair* | NATURAL OBJECT | *flower* | RELATION | *portion* |
| COGNITION | *way* | OTHER | *stuff* | SHAPE | *square* |
| COMMUNICATION | *review* | PERSON | *people* | STATE | *pain* |
| FEELING | *discomfort* | PHENOMENON | *result* | SUBSTANCE | *oil* |
| FOOD | *food* | | | TIME | *day* |

**Figure 19.2** Supersenses: 26 lexicographic categories for nouns in WordNet.

| Supersense | Verbs denoting ... |
|---|---|
| body | grooming, dressing and bodily care |
| change | size, temperature change, intensifying |
| cognition | thinking, judging, analyzing, doubting |
| communication | telling, asking, ordering, singing |
| competition | fighting, athletic activities |
| consumption | eating and drinking |
| contact | touching, hitting, tying, digging |
| creation | sewing, baking, painting, performing |
| emotion | feeling |
| motion | walking, flying, swimming |
| perception | seeing, hearing, feeling |
| possession | buying, selling, owning |
| social | political and social activities and events |
| stative | being, having, spatial relations |
| weather | raining, snowing, thawing, thundering |

# General Dictionaries

- Function words (e.g. for, rather, than), also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced “Luke”): Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. “emotion”, “cognition”, “work”, “family”, “positive”, “negative” etc.
- Mohammad and Turney (2011):
  - code 10,000 words along four emotional dimensions: joy–sadness, anger-fear, trust-disgust, anticipation-surprise
- Warriner et al (2013):
  - code 14,000 words along three emotional dimensions: valence, arousal, dominance.

# Lexicon-based Sentiment Analysis

- Extract a “tone” dimension – positive, negative, neutral
    - standard approach is lexicon-based, but they fail easily: e.g., “good” versus “not good” versus “not very good”
    - Off-the-shelf scores may be trained on biased corpora, eg online writing
    - Hamilton et al (2016) and Zorn and Rice (2019) show how to make domain-specific sentiment lexicons using word embeddings (more on this later).

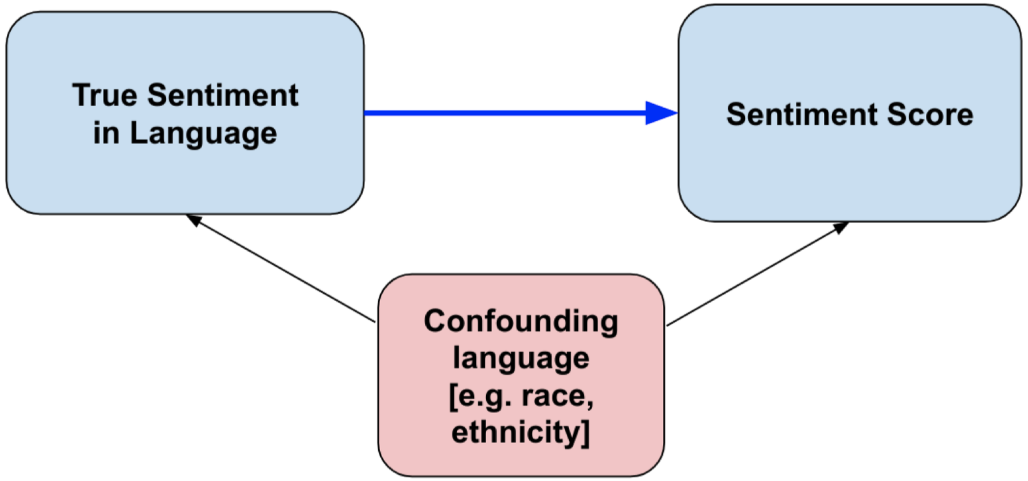
# Problems with Sentiment Analyzers: NLP System Bias

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

# NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.




- Supervised sentiment models are confounded by correlated language factors.
    - e.g., in the training set maybe people complain about Mexican food more often than Italian food because Italian restaurants tend to be more upscale.

# This is a universal problem

- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- dictionary methods, while having other limitations, mitigate this problem
  - the researcher intentionally “regularizes” out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.

# Supervised Classification

What is supervised classification?

- The learned prediction of the most likely of a set of $k > 1$ predefined nominal classes for an instance.

Learning phase (training)

- Input. A set of known instances $x^(i)$ with correct output class $c(x^{(i)})$.
- Output. A model $X \to C$ that maps any instance to its output class.

Application phase (prediction)

- Input. A set of unknown instances $x^(i)$ without output classes.
- Output. The output class $c(x^{(i)})$ for each instance.

# Feature-based Classification

Feature-based representation

- A feature vector is an ordered set of values of the form $x = (x_1, \ldots, x_m)$.
- Each feature $x_j$ denotes a measurable property of an input, $1 \leq j < m$.
- Each instance $o_j$ is mapped to a vector $x^{(i)} = (x_1^{(u)}, \ldots, x_m^{(i)})$ where $x_j^{(i)}$ denotes the value of feature $x_j$ .

Text mining using feature-based classification

- The main challenge is to engineer features that help solve a given task.
- In addition, a suitable classification algorithm needs to be chosen.

# Classification Algorithms

Binary vs. multiple-class classification (recap)

- Binary. Many classification algorithms work for $k = 2$ classes only.
- Multiple. Handled via multiple binary classifiers, e.g., one-versus-all.

Selected supervised classification algorithms

- Naïve Bayes. Predicts classes based on conditional probabilities.
- Support vector machine. Maximizes the margin between classes.
- Decision tree. Sequentially compares instances on single features.
- Random forest. Majority voting based on several decision trees.
- Neural network. Learns complex functions on feature combinations.
- ... and many more

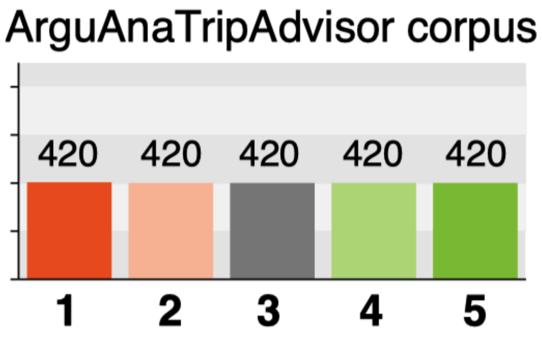
# Sentiment Analysis

**Sentiment classification of reviews**

- Classification of the nominal sentiment polarity or score of a customer review on a product, service, or work of art.

**Data**

- 2100 English hotel reviews from TripAdvisor.
  900 training, 600 validation, and 600 test reviews.
- Each review has a sentiment score from {1, ..., 5}.



# Sentiment classification of reviews

**Tasks**

- 3-class sentiment. 1–2 mapped to negative, 3 to neutral, 4–5 to positive. Training set balanced with random undersampling.
- 5-class sentiment. Each score interpreted as one (nominal) class.

**Approach**

- Algorithm. Linear SVM with one-versus-all multi-class handling.
- Features. Combination of several standard and specific feature types.

# Feature Engineering

**What is feature engineering?**

- The design and development of the feature representation of instances used to address a given task.
- The representation governs what patterns can be found during learning.

**Standard vs. specific features**

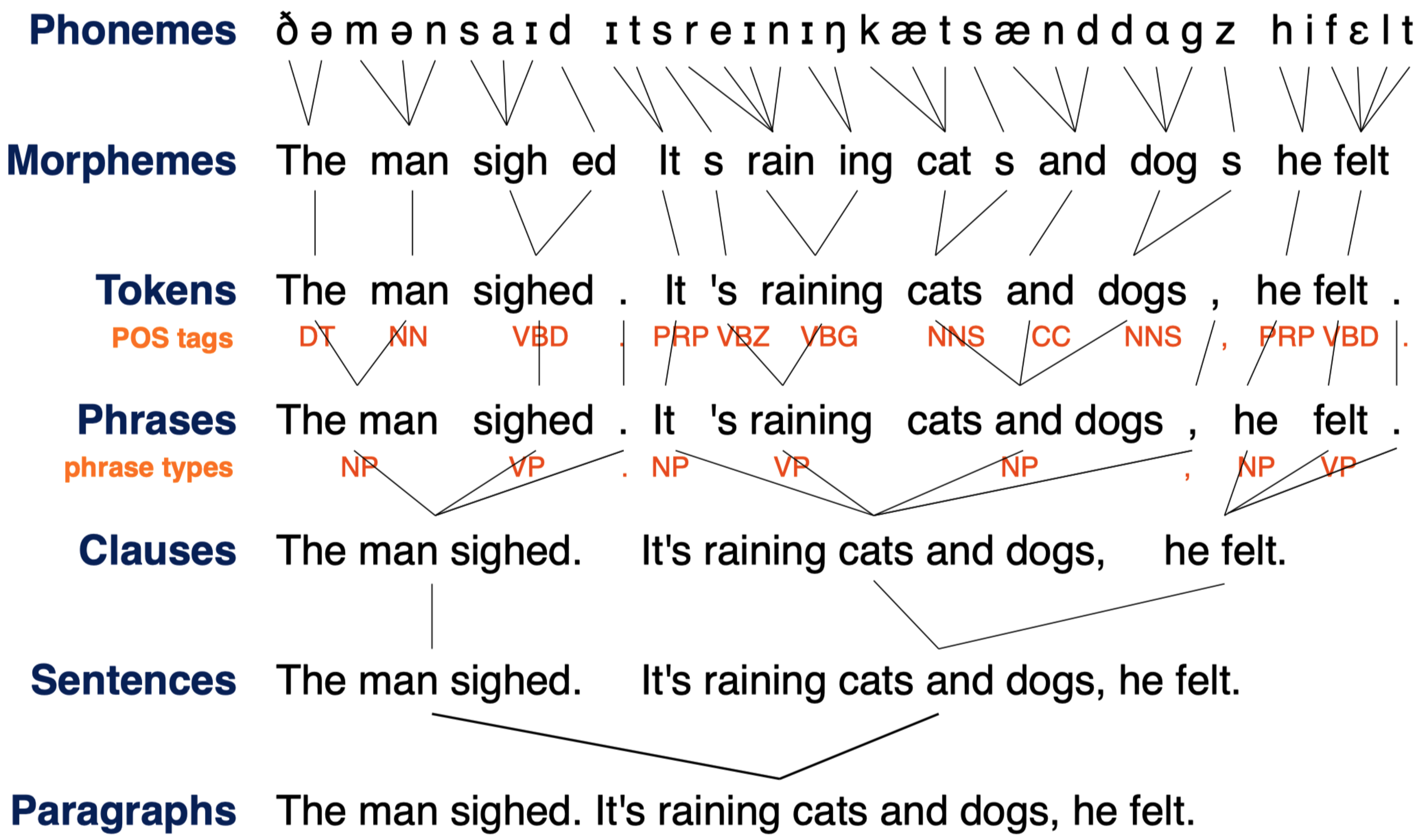
- Standard. Features that can be derived from (more or less) general linguistic phenomena and that may help in several tasks.
- Specific. Features that are engineered for a specific tasks, usually based on expert knowledge about the task.

# Feature Engineering

Features covered here

- Standard content features. Token n-grams, target class features.
- Standard style features. POS and phrase n-grams, stylometric features.
- Specific features. Local sentiment, discourse relations, flow patterns.

# Some General Linguistic Phenomena



# Standard Content Feature Types

**Token n-grams**

- Token unigrams (bag-of-words). The distribution of all token 1-grams that occur in at least 5% of all training texts.
- Token bigrams/trigrams. Analog for 2-grams and 3-grams.

**Target class features**

- Core vocabulary. The distribution of all words that occur at least three times as often in one class as in every other.
- Sentiment scores. The mean positivity, negativity, and objectivity of all first and average word senses in SentiWordNet.
- Sentiment words. The distribution of all subjective words in SentiWordNet.

# SentiWordNet

## 3. Visualizing SENTIWORDNET

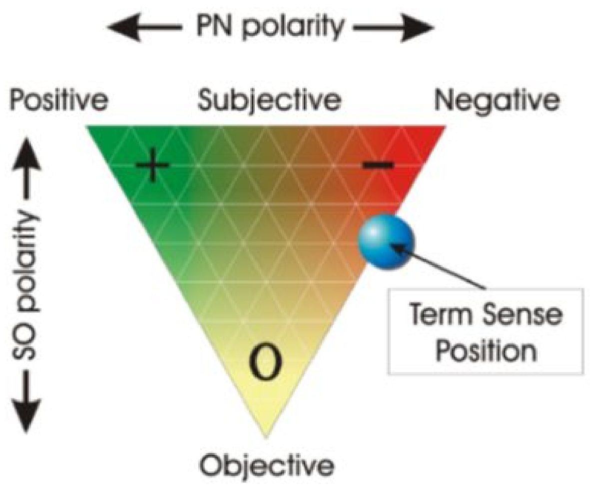



Figure 1: The graphical representation adopted by SENTIWORDNET for representing the opinion-related properties of a term sense.

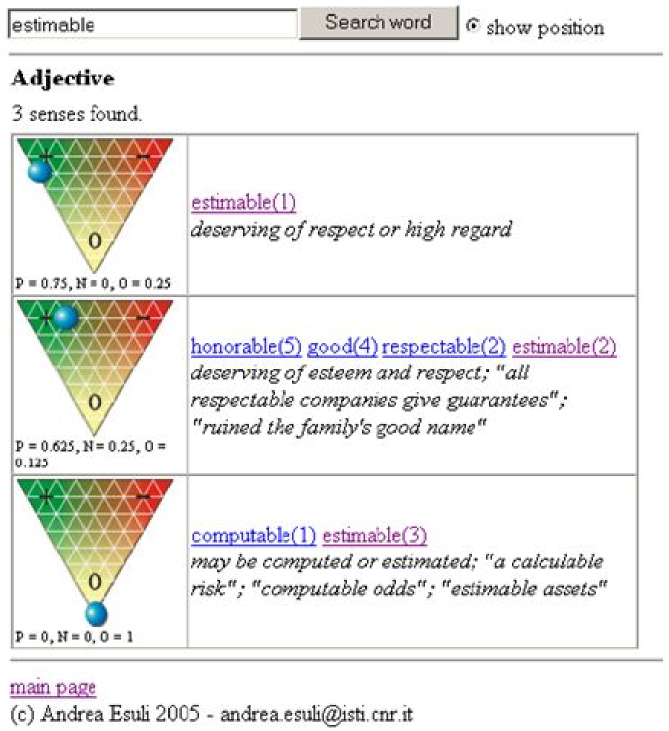



Figure 2: SENTIWORDNET visualization of the opinion-related properties of the term `estimable`.

# Standard Style Feature Types

**Part-of-speech (POS) tag n-grams**

- POS unigrams. The distribution of all part-of-speech 1-grams that occur in at least 5% of all training texts.
- POS bigrams/trigrams. Analog for 2-grams and 3-grams.

**Phrase type n-grams**

- Phrase unigrams. The distribution of all phrase type 1-grams that occur in at least 5% of all training texts.
- Phrase bigrams/trigrams. Analog for 2-grams and 3-grams.

# Standard Style Feature Types

**Stylometric features**

- Character trigrams. The distribution of all character 3-grams that occur in at least 5% of all training texts.
- Function words. The distribution of the top 100 words in the training set.
- Lexical statistics. Average numbers of tokens, clauses, and sentences.

# Evaluation of the Standard Feature Types

**Effectiveness results (accuracy)**

| Category | Feature type | # Features | 3 classes |
|---|---|---|---|
| Content | Token unigrams | 426 | **60.8%** |
| | Token bigrams | 112 | 49.5% |
| | Token trigrams | 64 | 24.5% |
| | Core vocabulary | 83 | 41.7% |
| | Sentiment scores | 6 | 59.3% |
| | Sentiment words | 123 | 60.5% |
| Style | POS unigrams | 48 | 51.3% |
| | POS bigrams | 70 | 49.0% |
| | POS trigrams | 118 | 45.5% |
| | Phrase unigrams | 13 | 48.8% |
| | Phrase bigrams | 43 | 52.5% |
| | Phrase trigrams | 122 | 50.8% |
| | Function words | 100 | 57.3% |
| | Character trigrams | 200 | 48.7% |
| | Lexical statistics | 6 | 42.8% |
| **Combination of features** | | **1534** | **60.8%** |

# Evaluation of the Standard Feature Types

### Evaluation

- One linear SVM for each feature type alone and for their combination.
- Training on training set, tuning on validation set, test on test set.

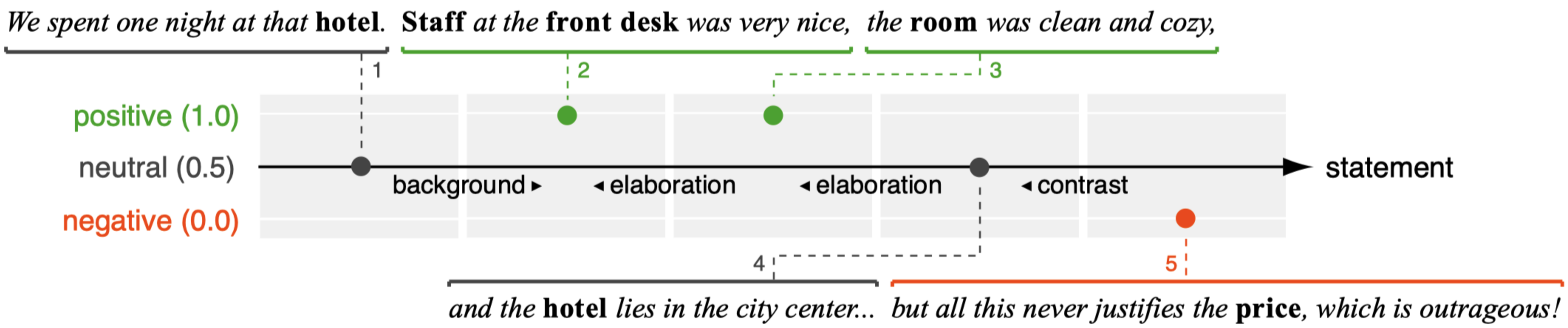
### Discussion

- Token unigrams best, but some other types close.
- Combination does not outperform single types.
- 60.8% accuracy does not seem satisfying.

# Review Argumentation

### Example hotel review

"We spent one night at that hotel. Staff at the front desk was very nice, the room was clean and cozy, and the hotel lies in the city center... but all this never justifies the price, which is outrageous!"



# Review Argumentation

**A shallow model of review argumentation**

- A review can be seen as a flow of local sentiments on domain concepts that are connected by discourse relations.

# Specific Feature Types for Review Sentiment Analysis

**Local sentiment distribution**

- The frequencies of positive, neutral, and negative local sentiment as well as of changes of local sentiments.

  > positive 0.4 neutral 0.4 negative 0.2 (neutral, positive) 0.25 ...

- The average local sentiment value from 0.0 (negative) to 1.0 (positive).

  > average sentiment 0.6

- The interpolated local sentiment at each normalized position in the text.

  > e.g., normalization length 9: (0.5, 0.75, 1.0, 1.0, 1.0, 0.75, 0.5, 0.25, 0.0)

# Specific Feature Types for Review Sentiment Analysis

**Discourse relation distribution**

- The distribution of discourse relation types in the text.

> background 0.25 elaboration 0.5 contrast 0.25 (all others 0.0)

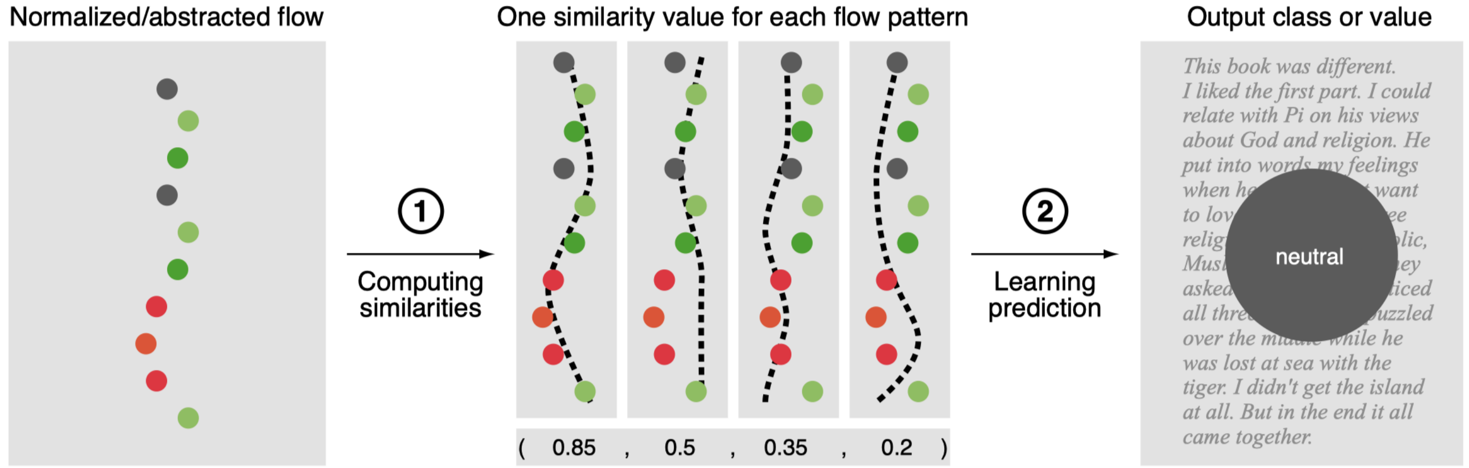
- The distribution of combinations of relation types and local sentiments.

> background(neutral, positive) 0.25 elaboration(positive, positive) 0.25 ...

# Specific Feature Types for Review Sentiment Analysis

## Sentiment flow patterns

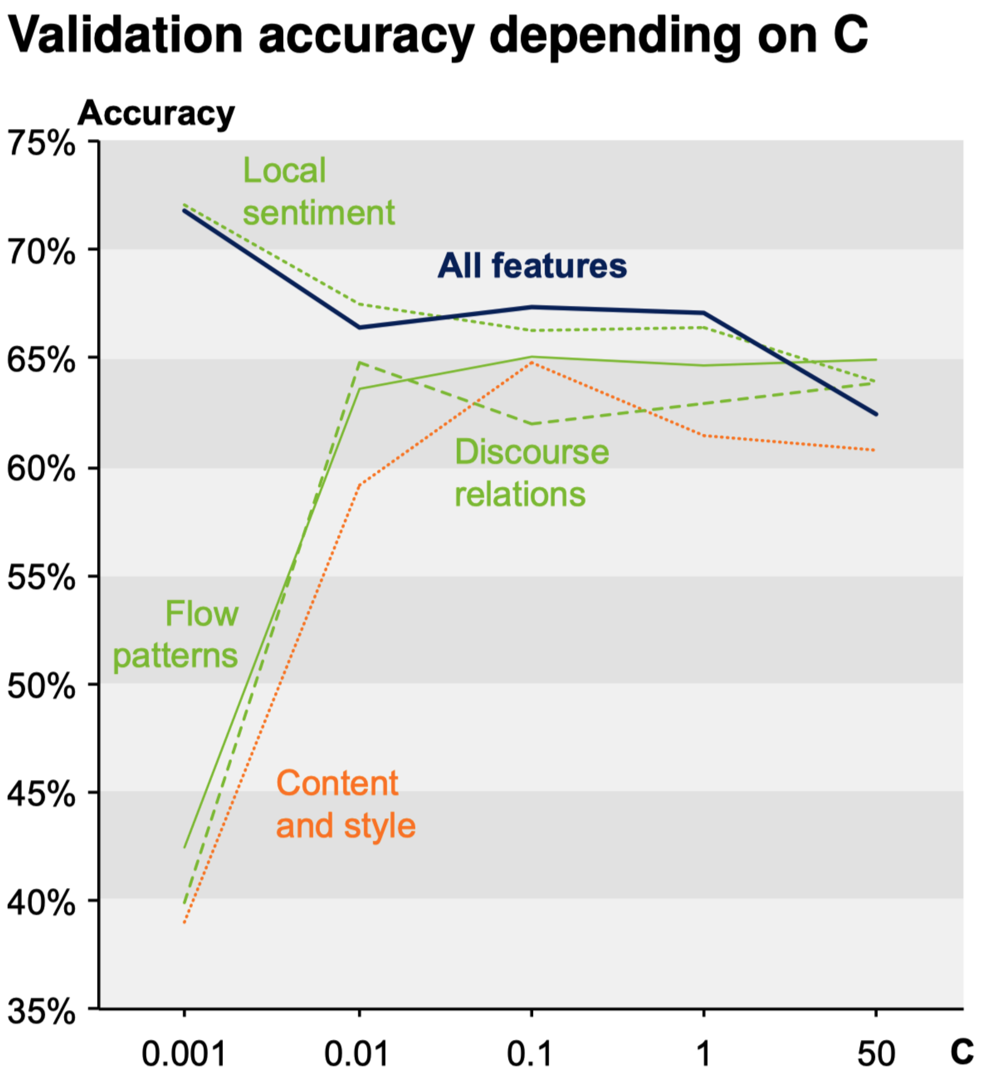
- The similarity of the normalized flow of the text to each flow pattern.




## Content and style features

- Content. Token n-grams, sentiment scores.
- Style. Part-of-speech n-grams, character trigrams, lexical statistics.

# Evaluation of the Specific Feature Types



# Evaluation of the Specific Feature Types

## Evaluation

- One linear SVM for each feature type alone and for their combination.
- Training on training set, tuning on validation set, test on test set.
- Both 3-class and 5-class.

## Cost hyperparameter tuning

- Tested $C$ values. 0.001, 0.01, 0.1, 1.0, 50.0
- Best $C$ used on test set.
- Results shown here for the 3-class task only.

# Results and Discussion for the Specific Features

Effectiveness results on test set (accuracy)

| **Feature type** | **# Features** | **3 Classes** | **5 Classes** |
|---|---|---|---|
| Local sentiment distribution | 50 | 69.8% | 42.2% |
| Discourse relation distribution | 75 | 65.3% | 40.6% |
| Sentiment flow patterns | 42 | 63.1% | 39.7% |
| Content and style features | 1026 | 58.9% | 43.2% |
| **Combination of features** | 1193 | **71.5%** | **48.1%** |
| Random baseline | | 33.3% | 20.0% |

# Results and Discussion for the Specific Features

**Discussion**

- Content and style features. A bit weaker than in the experiment above, due to slight differences in the experiment setting.
- Sentiment flow patterns. Impact is more visible across domains.
- Combination of features. Works out this time, so more complementary.
- The 5-class accuracy seems insufficient.
- Classification misses to model the ordinal relation between classes; regression might be better.