

Natural Language Processing 1

Katia Shutova

ILLC
University of Amsterdam

29 October 2016

Lecture 1: Introduction

Lecture 1: Introduction

Overview of the course

NLP applications

Why NLP is hard

Sentiment classification

Overview of the practical

Taught by...



Katia Shutova
Lecturer
e.shutova@uva.nl



Joost Bastings
Lab & practical
coordinator
j.bastings@uva.nl



Samira Abnار
Senior TA
s.abnar@uva.nl

Teaching assistants



Daniel Daza



Victor Milewski



Laura Ruis



Jack Harding



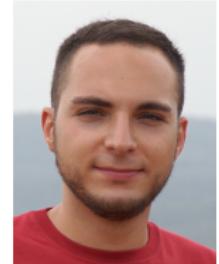
Mattijs Mul



Florian Mohnert



Jaap Jumelet



Mario Giulianelli

Overview of the course

- ▶ Introduction and broad overview of NLP
- ▶ Different levels of language analysis (word, sentence, larger text fragments)
- ▶ A range of NLP tasks and applications
- ▶ Both fundamental and most recent methods:
 - ▶ rule-based
 - ▶ statistical
 - ▶ deep learning
- ▶ Other NLP courses go into much greater depth

Assessment

1. Practical assignments (50%)

- ▶ Work in groups of 2
- ▶ Implement several language processing methods
- ▶ Evaluate in the context of a real-world NLP application — sentiment classification
- ▶ Assessed by two reports (25% each)
 - ▶ Practical 1: Mid-term report, deadline **23 November**
 - ▶ Practical 2: Final report, deadline **12 December**

2. Exam on 21 December (50%)

- ▶ Exam preparation exercises (individual work)
- ▶ feedback from TAs

Need to pass **both** components to get a passing grade

Also note:

Course materials and more info:

<https://cl-illc.github.io/nlp1/>

Contact

- ▶ Main contact – your TA (email on the website)
- ▶ Katia: e.shutova@uva.nl
- ▶ Joost: j.bastings@uva.nl

Subject line should have **NLP1-18**

Email your TA by Weds, 31 October with details of your group.

- ▶ names of the students
- ▶ their email addresses

Course Materials

- ▶ Slides, further reading, assignments posted on the **website**
- ▶ **but...** assignment submission will be via **Canvas**.
- ▶ **Book:** Jurafsky & Martin, *Speech and Language Processing (2nd edition)*
3 edition (unofficial) at
<https://web.stanford.edu/~jurafsky/slp3/>
- ▶ For most topics, additional (optional) readings of **research papers** put up on the website.

What is NLP?

NLP: the computational modelling of human language.

Many popular applications



...and the emerging ones



Machine Translation

- ▶ Translate from one language into another
- ▶ Earliest attempted NLP application
- ▶ High quality with typologically close languages: e.g. Swedish-Danish.
- ▶ More challenging with typologically distant languages and low-resource languages
- ▶ Early systems based on transfer rules, then statistical and now neural MT

Retrieving information

- ▶ **Information retrieval:** return documents in response to a user query (Internet Search is a special case)
- ▶ **Information extraction:** discover specific information from a set of documents (e.g. companies and their founders)
- ▶ **Question answering:** answer a specific user question by returning a section of a document:

What is the capital of France?

Paris has been the French capital for many centuries.

Opinion mining and sentiment analysis

- ▶ Finding out what people think about politicians, products, companies etc.
- ▶ Increasingly done on web documents and social media
- ▶ More about this later today



Emerging applications

Automated fact checking

- ▶ classify statements and news articles as factual or not
- ▶ in an effort to combat disinformation



Abusive language detection

- ▶ automated detection and moderation of online abuse
- ▶ hate speech, racism, sexism, personal attacks, cyberbullying etc.



Other areas in which NLP is relevant

NLP and computer vision

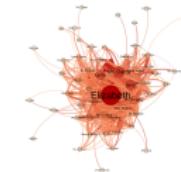
- ▶ Caption generation for images and videos



The dog chewed at the shoes

Digital humanities

- ▶ e.g. social network in *Pride and Prejudice*



Computational social science

- ▶ analyse human behaviour based on language use (deeper than sentiment)



NLP and linguistics

1. **Morphology** — the structure of words: lecture 2.
2. **Syntax** — the way words are used to form phrases: lectures 3 and 4.
3. **Semantics**
 - ▶ **Lexical semantics** — the meaning of individual words: lectures 5 and 6.
 - ▶ **Compositional semantics** — the construction of meaning of longer phrases and sentences (based on syntax): lectures 7 and 9.
4. **Pragmatics** — meaning in context: lectures 8 and 10.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- ▶ Word senses: bank (finance or river?)
- ▶ Part of speech: chair (noun or verb?)
- ▶ Syntactic structure: I saw a man with a telescope
- ▶ Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- ▶ Word senses: **bank** (finance or river?)
- ▶ Part of speech: **chair** (noun or verb?)
- ▶ Syntactic structure: I saw a man with a telescope
- ▶ Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- ▶ Word senses: bank (finance or river?)
- ▶ Part of speech: chair (noun or verb?)
- ▶ Syntactic structure: I saw a man with a telescope
- ▶ Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- ▶ Word senses: bank (finance or river?)
- ▶ Part of speech: chair (noun or verb?)
- ▶ Syntactic structure: I saw a man with a telescope
- ▶ Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Why is NLP hard?

Ambiguity: *same strings can mean different things*

- ▶ Word senses: bank (finance or river?)
- ▶ Part of speech: chair (noun or verb?)
- ▶ Syntactic structure: I saw a man with a telescope
- ▶ Multiple: I saw her duck

Finally, a computer that understands you like your mother!

Ambiguity grows with sentence length, sometimes exponentially.

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Real examples from newspaper headlines

Iraqi head seeks arms

Stolen painting found by tree

Teacher strikes idle kids

Why is NLP hard?

Synonymy and variability: different strings can mean the same or similar things

Did Google buy YouTube?

1. Google purchased YouTube
2. Google's acquisition of YouTube
3. Google acquired every company
4. YouTube may be sold to Google
5. Google didn't take over YouTube

Wouldn't it be better if ... ?

The properties which make natural language difficult to process are essential to human communication:

- ▶ Flexible
- ▶ Learnable, but expressive and compact
- ▶ Emergent, evolving systems

Synonymy and ambiguity go along with these properties.

Natural language communication can be indefinitely precise:

- ▶ Ambiguity is mostly local (for humans)
- ▶ resolved by immediate context
- ▶ but requires world knowledge

Wouldn't it be better if ... ?

The properties which make natural language difficult to process are essential to human communication:

- ▶ Flexible
- ▶ Learnable, but expressive and compact
- ▶ Emergent, evolving systems

Synonymy and ambiguity go along with these properties.

Natural language communication can be indefinitely precise:

- ▶ Ambiguity is mostly local (for humans)
- ▶ resolved by immediate context
- ▶ but requires world knowledge

World knowledge...

“Knowledge is knowing that a tomato is a fruit”



BUT



“Wisdom is knowing not to put it in a fruit salad”

- ▶ Impossible to hand-code at a large-scale
- ▶ either limited domain applications
- ▶ or learn approximations from the data

Opinion mining: what do they think about me?

- ▶ Task: scan documents (webpages, tweets etc) for positive and negative opinions on people, products etc.
- ▶ Find all references to entity in some document collection: list as positive, negative (possibly with strength) or neutral.
- ▶ Construct summary report plus examples (text snippets).
- ▶ Fine-grained classification:
e.g., for phone, opinions about: overall design, display, camera.

LG G3 review (Guardian 27/8/2014)

The shiny, brushed effect makes the G3's plastic design looks deceptively like metal. It feels solid in the hand and the build quality is great — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

*The shiny, brushed effect makes the G3's plastic **design** looks deceptively like metal. It feels solid in the hand and the build quality is great — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

*The shiny, brushed effect makes the G3's plastic **design** looks deceptively like metal. It feels solid in the hand and the **build quality** is **great** — there's minimal give or flex in the body. It weighs 149g, which is lighter than the 160g HTC One M8, but heavier than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

*The shiny, brushed effect makes the G3's plastic **design** looks deceptively like metal. It feels solid in the hand and the **build quality** is great — there's minimal give or flex in the body. It **weighs** 149g, which is **lighter** than the 160g HTC One M8, but **heavier** than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

The G3's claim to fame is its 5.5in quad HD display, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.

LG G3 review (Guardian 27/8/2014)

*The shiny, brushed effect makes the G3's plastic **design** looks deceptively like metal. It feels solid in the hand and the **build quality** is great — there's minimal give or flex in the body. It **weighs** 149g, which is **lighter** than the 160g HTC One M8, but **heavier** than the 145g Galaxy S5 and the significantly smaller 112g iPhone 5S.*

*The G3's **claim to fame** is its 5.5in quad HD **display**, which at 2560x1440 resolution has a pixel density of 534 pixels per inch, far exceeding the 432ppi of the Galaxy S5 and similar rivals. The screen is vibrant and crisp with wide viewing angles, but the extra pixel density is not noticeable in general use compared to, say, a Galaxy S5.*

Sentiment classification: the research task

- ▶ Full task: information retrieval, cleaning up text structure, named entity recognition, identification of relevant parts of text. Evaluation by humans.
- ▶ Research task: preclassified documents, topic known, opinion in text along with some straightforwardly extractable score.
- ▶ Pang et al. 2002: *Thumbs up? Sentiment Classification using Machine Learning Techniques*
- ▶ Movie review **corpus**: strongly positive or negative reviews from IMDb, 50:50 split, with rating score.

Sentiment analysis as a text classification problem

- ▶ *Input:*
 - ▶ a document d
 - ▶ a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
- ▶ *Output:*
 - ▶ a predicted class $c \in C$

IMDb: An American Werewolf in London (1981)

Rating: 9/10

Ooooo. Scary.

The old adage of the simplest ideas being the best is once again demonstrated in this, one of the most entertaining films of the early 80's, and almost certainly Jon Landis' best work to date. The script is light and witty, the visuals are great and the atmosphere is top class. Plus there are some great freeze-frame moments to enjoy again and again. Not forgetting, of course, the great transformation scene which still impresses to this day.

In Summary: Top banana

Bag of words representation

Treat the reviews as collections of individual words.

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



Bag of words representation

- ▶ Classify reviews according to positive or negative words.
- ▶ Could use word lists prepared by humans — **sentiment lexicons**
- ▶ but machine learning based on a portion of the corpus (**training set**) is preferable.
- ▶ Use human rankings for training and evaluation.

Supervised classification

- ▶ *Input:*
 - ▶ a document d
 - ▶ a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
 - ▶ a training set of m hand-labeled documents
 $(d_1, c_1), \dots, (d_m, c_m)$
- ▶ *Output:*
 - ▶ a learned classifier $\gamma : d \rightarrow c$

Classification methods

Many classification methods available

- ▶ Naive Bayes
- ▶ Logistic regression
- ▶ Decision trees
- ▶ k-nearest neighbors
- ▶ Support vector machines
- ▶ ...

Documents as feature vectors

The document d is represented as a **feature vector** \vec{f} :

I love this movie! It's sweet,
but with satirical humor. The
dialogue is great and the
adventure scenes are fun...
It manages to be whimsical
and romantic while laughing
at the conventions of the
fairy tale genre. I would
recommend it to just about
anyone. I've seen it several
times, and I'm always happy
to see it again whenever I
have a friend who hasn't
seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Naive Bayes classifier

Choose most probable class given a feature vector \vec{f} :

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|\vec{f})$$

Apply Bayes Theorem:

$$P(c|\vec{f}) = \frac{P(\vec{f}|c)P(c)}{P(\vec{f})}$$

Constant denominator:

$$\hat{c} = \operatorname{argmax}_{c \in C} P(\vec{f}|c)P(c)$$

Naive Bayes: feature independence

$$\hat{c} = \operatorname{argmax}_{c \in C} P(\vec{f}|c)P(c)$$

Problem: need a very, very large corpus to estimate $P(\vec{f}|c)$

$$P(\vec{f}|c) = P(f_1, f_2, \dots, f_n|c)$$

Conditional independence assumption ('naive'): assume the feature probabilities $P(f_i|c)$ are independent given the class c .

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) \prod_{i=1}^n P(f_i|c)$$

Naive Bayes: feature independence

$$\hat{c} = \operatorname{argmax}_{c \in C} P(\vec{f}|c)P(c)$$

Problem: need a very, very large corpus to estimate $P(\vec{f}|c)$

$$P(\vec{f}|c) = P(f_1, f_2, \dots, f_n|c)$$

Conditional independence assumption ('naive'): assume the feature probabilities $P(f_i|c)$ are independent given the class c .

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) \prod_{i=1}^n P(f_i|c)$$

Naive Bayes: Learning the model

Maximum likelihood estimation: use frequencies in the data

$$\hat{P}(c) = \frac{\text{Doccount}(c)}{N_{doc}}$$

$$\hat{P}(f_i|c) = \frac{\text{count}(f_i, c)}{\sum_{f \in V} \text{count}(f, c)}$$

Problem with maximum likelihood

What if we have seen no training documents with the word ***fantastic*** and classified as **positive**?

$$\hat{P}(\text{fantastic}|\text{positive}) = \frac{\text{count}(\text{fantastic}, \text{positive})}{\sum_{f \in V} \text{count}(f, \text{positive})} = 0$$

Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) \prod_{i=1}^n P(f_i|c)$$

Laplace smoothing for Naive Bayes

Smoothing is a way to handle data sparsity

Laplace (also called "add 1") smoothing:

$$\hat{P}(f_i|c) = \frac{\text{count}(f_i, c) + 1}{\sum_{f \in V} (\text{count}(f, c) + 1)} = \frac{\text{count}(f_i, c) + 1}{\sum_{f \in V} \text{count}(f, c) + |V|}$$

Log space

Use **log space** to prevent arithmetic underflow.

- ▶ Multiplying lots of probabilities can result in floating-point underflow
- ▶ sum logs of probabilities instead of multiplying probabilities

$$\log(xy) = \log(x) + \log(y)$$

- ▶ class with the highest log probability score is still the most probable

$$\hat{c} = \operatorname{argmax}_{c \in C} (\log P(c) + \sum_{i=1}^n \log P(f_i|c))$$

Test sets and cross-validation

Divide the corpus into

- ▶ **training** set — to train the model
- ▶ **development** set — to optimize its parameters
- ▶ **test** set — kept unseen to avoid overfitting

or...

use **cross-validation** over multiple splits

- ▶ divide the corpus into e.g. 10 parts
- ▶ train on 9 parts, test on 1 part
- ▶ average results from all splits

Evaluation

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

Pang et al. (2002):

- ▶ The corpus is artificially balanced
- ▶ Chance success is 50%
- ▶ Bag-of-words achieves an accuracy of 80%.

Precision and Recall

What if the corpus were not balanced?

- ▶ **Precision:** % of selected items that are correct
- ▶ **Recall:** % of correct items that are selected

		<i>gold standard labels</i>		
		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$
	system negative	false negative	true negative	
		$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$		$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}}$

F-measure

Also called *F-score*

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2P + R}$$

β controls the importance of recall and precision

$\beta = 1$ is typically used:

$$F_1 = \frac{2PR}{P + R}$$

Error analysis

Bag-of-words gives **80% accuracy** in sentiment analysis

Some sources of **errors**:

- ▶ Negation:

Ridley Scott has never directed a bad film.

- ▶ Overfitting the training data:

e.g., if training set includes a lot of films from before 2005,
Ridley may be a strong positive indicator, but then we test
on reviews for ‘Kingdom of Heaven’?

- ▶ Comparisons and contrasts.

Contrasts in the discourse

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

More contrasts

AN AMERICAN WEREWOLF IN PARIS is a failed attempt . . . Julie Delpy is far too good for this movie. She imbues Serafine with spirit, spunk, and humanity. This isn't necessarily a good thing, since it prevents us from relaxing and enjoying AN AMERICAN WEREWOLF IN PARIS as a completely mindless, campy entertainment experience. Delpy's injection of class into an otherwise classless production raises the specter of what this film could have been with a better script and a better cast . . . She was radiant, charismatic, and effective . . .

Doing sentiment classification ‘properly’?

- ▶ Morphology, syntax and compositional semantics:
who is talking about what, what terms are associated with
what, tense ...
- ▶ Lexical semantics:
are words positive or negative **in this context**? Word
senses (e.g., *spirit*)?
- ▶ Pragmatics and discourse structure:
what is the topic of this section of text? Pronouns and
definite references.
- ▶ Getting all this to work well on arbitrary text is very hard.
- ▶ Ultimately the problem is **AI-complete**, but can we do well
enough for NLP to be useful?

Human translation?



Human translation?



I am not in the office at the moment. Please send any work to be translated.

Sentiment analysis practical: Part 1

Sentiment classification of movie reviews

1. Sentiment classification with a **sentiment lexicon**
2. Implement **Naive Bayes** classifier with **bag-of-word** features
3. Model **grammar**: word order and part of speech tags
4. Experiment with **support vector machines** (SVM) classifier
5. Evaluate and compare different methods

Assessed by the **mid-term report**, deadline 23 November

Sentiment analysis practical: Part 1

Sentiment classification of movie reviews

1. Sentiment classification with a **sentiment lexicon**
2. Implement **Naive Bayes** classifier with **bag-of-word** features
3. Model **grammar**: word order and part of speech tags
4. Experiment with **support vector machines** (SVM) classifier
5. Evaluate and compare different methods

Assessed by the **mid-term report**, deadline 23 November

Sentiment analysis practical: Part 2

- ▶ Experiment within a **deep learning** framework
- ▶ Include **semantics**
- ▶ Model the meaning of words, phrases and sentences
- ▶ Evaluate in the sentiment classification task

Assessed by the **final report**, deadline 12 December

Sentiment analysis practical: Part 2

- ▶ Experiment within a **deep learning** framework
- ▶ Include **semantics**
- ▶ Model the meaning of words, phrases and sentences
- ▶ Evaluate in the sentiment classification task

Assessed by the **final report**, deadline 12 December