Content and Style Analysis

Die Pandemie in Sprache und Text – Corona-Podcasts & co.

Jürgen Hermes, Lukas Mönch, Felix Rau, Nils Reiter

May 7, 2021

Today

Stylometry

Sentiment Analysis

Finding Text Pieces

Section 1

Stylometry

Introduction

- Application: Authorship attribution/verification, plagiarism detection/obfuscation
- Scholarly use case: Author style

Introduction

- Application: Authorship attribution/verification, plagiarism detection/obfuscation
- Scholarly use case: Author style
- Stylometry: Original branch of digital humanities

Stylometry

Examples

- Federalist Papers (Mosteller and Wallace, 1964) - based on frequency of function words (prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles), rather than content words.
- J.K. Rowling and “Robert Galbraith” (Juola, 2013)
- Dutch national anthem (Kestemont et al., 2017)
- Ellena Ferrante novels (Savoy, 2018)
- Multiple authors in one text:
  - The Book of Mormon (Jockers, 2008; Schaalje et al., 2011)
  - “Collaborative authorship: Conrad, Ford and Rolling Delta” (Rybicki, Hoover, Kestemont, 2014)

Observation

- Authorship attribution works well using most frequent words
- Surprising: Most frequent words are function words
  - Remember Zipf's law

Observation

- Authorship attribution works well using most frequent words
- Surprising: Most frequent words are function words
  - Remember Zipf's law
- Intuition: Function words are used unconsciously
  - Guided by the way we formulate and syntactic rules

Classification vs. Clustering

- Classification approach: Given this text, which of these people (= classes) is the author?
  - Solved problem, but no generalisation: Every text and set of people needs a new model

Classification vs. Clustering

- Classification approach: Given this text, which of these people (= classes) is the author?
  - Solved problem, but no generalisation: Every text and set of people needs a new model
- Classification approach: Are these two texts written by the same author (classes: yes/no)
  - Highly skewed distribution: for any set of texts, most pairs are not written by the same author

Classification vs. Clustering

- Classification approach: Given this text, which of these people (= classes) is the author?
  - Solved problem, but no generalisation: Every text and set of people needs a new model
- Classification approach: Are these two texts written by the same author (classes: yes/no)
  - Highly skewed distribution: for any set of texts, most pairs are not written by the same author
- Clustering approach: Given these texts, group them according to likely authors
  - Most productive branch of stylometry ✔

Clustering Texts

- Clustering is done on features, just as classification
- Which features do we use? Word frequencies

Clustering Texts

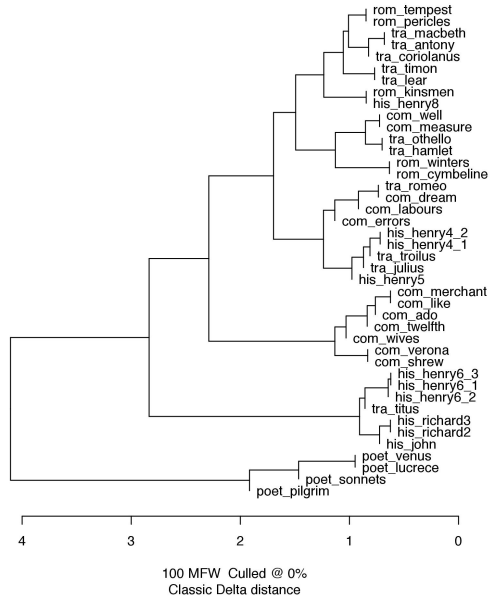
- Clustering is done on features, just as classification
- Which features do we use? Word frequencies
- Workflow
  1. Extract document-term-matrix, represent each document with one vector
     - Optional: Restrict to most frequent/important/interesting terms
  2. Calculate distance between all documents
  3. Group them according to distance

Distance Functions

- Cosine similarity
- Euclidean distance
- Manhatten distance
- Burrow's Delta
- ...

Stylo R package

- https://computationalstylistics.github.io
- Eder, Rybicki, Kestemont
- Tutorials
- Many statistics, options and outputs included
- DH Summer School @ University of Leipzig teaches it most years, 3rd to 13th of August 2021, https://esu.culintec.de


my first experiment
Cluster Analysis
rom_tempest
rom_pericles
tra_macbeth
tra_antony
tra_coriolanus
tra_timon
tra_lear
rom_kinsmen
his_henry8
com_well
com_measure
tra_othello
tra_hamlet
rom_winters
rom_cymbeline
tra_romeo
com_dream
com_labours
com_errors
his_henry4_2
his_henry4_1
tra_troilus
tra_julius
his_henry5
com_merchant
com_like
com_ado
com_twelfth
com_wives
com_verona
com_shrew
his_henry6_3
his_henry6_1
his_henry6_2
tra_titus
his_richard3
his_richard2
his_john
poet_venus
poet_lucrece
poet_sonnets
poet_pilgrim
4
3
2
1
0
100 MFW Culled @ 0%
Classic Delta distance

my first experiment
Bootstrap Consensus Tree

100-1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.75

demo

Section 2

Sentiment Analysis

Introduction

- Popular NLP area
- Sentiment: „Polarity“ of a text, i.e., positive or negative
  - Academic use cases: Movie or book reviews
  - Real-world application: Opinion mining, social media opinions on products

Introduction

- Popular NLP area
- Sentiment: „Polarity“ of a text, i.e., positive or negative
  - Academic use cases: Movie or book reviews
  - Real-world application: Opinion mining, social media opinions on products
- Fuzzy concept when taken out of its original context
  - What's a positive opinion about politics?

Introduction

- Popular NLP area
- Sentiment: „Polarity“ of a text, i.e., positive or negative
  - Academic use cases: Movie or book reviews
  - Real-world application: Opinion mining, social media opinions on products
- Fuzzy concept when taken out of its original context
  - What's a positive opinion about politics?
- Ilias: Slide sets from two sentiment analysis courses

Different Approaches

Text Classification

Typically short texts are classified in (two) classes

Different Approaches

Text Classification

Typically short texts are classified in (two) classes

Sentence classification

Each sentence is classified in (two) classes

Different Approaches

Text Classification

Typically short texts are classified in (two) classes

Sentence classification

Each sentence is classified in (two) classes

Aspect-oriented sentiment analysis

Given a sentence, which aspect of the target object is evaluated and how?

Different Approaches

Text Classification

Typically short texts are classified in (two) classes

Sentence classification

Each sentence is classified in (two) classes

Aspect-oriented sentiment analysis

Given a sentence, which aspect of the target object is evaluated and how?

- Different input representations and/or preprocessing requirements

Features

(Berendt, 2015)

- Words (bag-of-words)
- n-grams
- Parts of speech (e.g. Adjectives and adjective-adverb combinations)
- Opinion words (lexicon/dictionary-based or corpus)
- Valence intensifiers and shifters (for negation); modal verbs; ...
- Syntactic dependency

Feature Selection and Weighting

(Berendt, 2015)

Feature selection based on

- Frequency
- Information gain (Manning/Schütze, 1999, 583 f.)
- Odds ratio (for binary-class models; $OR(A,B) = \frac{\frac{f(A)}{1-f(A)}}{\frac{f(B)}{1-f(B)}}$) (JM19, 406 f.)
- Mutual information (Manning/Schütze, 1999, 66 f.)

Feature Selection and Weighting

(Berendt, 2015)

Feature selection based on

- Frequency
- Information gain (Manning/Schütze, 1999, 583 f.)
- Odds ratio (for binary-class models; $OR(A,B) = \frac{\frac{f(A)}{1-f(A)}}{\frac{f(B)}{1-f(B)}}$) (JM19, 406 f.)
- Mutual information (Manning/Schütze, 1999, 66 f.)

Feature weighting

- Term presence or term frequency
- Inverse document frequency (tf-idf) (Manning/Schütze, 1999, 542 ff.)
- Term position: e.g. title, first and last sentence(s)

ML Models

- Naive Bayes
- Logistic regression / maximum entropy
- Decision trees
- Support vector machines
- ...

Lexicons

- NRC Word-Emotion Association Lexicon Mohammad:2013aa
  - https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- Bing Liu's opinion lexicon
  - https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- MPQA subjectivity lexicon
  - http://www.cs.pitt.edu/mpqa/
- SentiWordNet
  - https://github.com/aesuli/SentiWordNet
- SenticNet
  - http://sentic.net

Annotated Data Sets

- Stanford sentiment treebank: `http://nlp.stanford.edu/sentiment/`
- Data from Lillian Lee's group: `http://www.cs.cornell.edu/home/llee/data/`
- Data from Bing Liu: `http://www.cs.uic.edu/~liub/`
- Large movie review dataset: `http://ai.stanford.edu/~amaas/data/sentiment/`
- Pranav Anand & co. (`http://people.ucsc.edu/~panand/data.php`):
  - Internet Argument Corpus
  - Annotated political TV ads
  - Focus of negation corpus
  - Persuasion corpus (blogs)

Resources for German

- Interest Group on German Sentiment Analysis (IGGSA)
  - `https://sites.google.com/site/iggsahome`
- General NLP (e.g., for preprocessing)
  - Spacy: `https://spacy.io` (Python)
  - DKpro: `https://dkpro.github.io/dkpro-core/` (Java/UIMA)

Tools

- SentiStrength `http://sentistrength.wlv.ac.uk` (Windows and Java)
- TheySay (`http://apidemo.theysay.io`) (REST API)
- Stanford Sentiment Analysis `http://nlp.stanford.edu/sentiment` (Java)
- VADER (`https://github.com/cjhutto/vaderSentiment`) (Python)

Section 3

Finding Text Pieces

Introduction

- Use NLP as a collection of methods and building blocks
- NLP: Linguistic tasks
- DH/CLS: Non-linguistic tasks
- Apply NLP building blocks to other things
- Identify most technically similar NLP task

Introduction

- Use NLP as a collection of methods and building blocks
- NLP: Linguistic tasks
- DH/CLS: Non-linguistic tasks
- Apply NLP building blocks to other things
- Identify most technically similar NLP task
- Task
  - Relatively few instances (e.g., five in a podcast episode)
  - An instance may consist of multiple words

Introduction

- Use NLP as a collection of methods and building blocks
- NLP: Linguistic tasks
- DH/CLS: Non-linguistic tasks
- Apply NLP building blocks to other things
- Identify most technically similar NLP task
- Task
  - Relatively few instances (e.g., five in a podcast episode)
  - An instance may consist of multiple words
- Similar NLP task: Named entity recognition

Named Entity Recognition

- Better name: Proper name detection
  - Because it's not about „entities that have names“, but about the names themselves

Named Entity Recognition

- Better name: Proper name detection
  - Because it's not about „entities that have names“, but about the names themselves

Examples

- The head of the European Commission, Ursula von der Leyen, has said the bloc is „ready to discuss“ a US-backed proposal for a waiver on the patents for Covid-19 vaccines [...].
  theguardian.com

Named Entity Recognition

- Better name: Proper name detection
  - Because it's not about „entities that have names", but about the names themselves

Examples

- The head of the European Commission, Ursula von der Leyen, has said the bloc is „ready to discuss" a US-backed proposal for a waiver on the patents for Covid-19 vaccines [...].
  theguardian.com
- Wie die Nachrichtenagentur Reuters unter Berufung auf die französische Journalistin Marie Carof-Gadel berichtet, fuhren französische Fischer am Donnerstagmorgen in den Hafen von Saint Helier auf der britischen Insel Jersey ein und positionierten ihr Boot dort unter anderem vor der Fähre „Commodore Goodwill". spiegel.de

Named Entity Recognition

- Better name: Proper name detection
  - Because it's not about „entities that have names", but about the names themselves

Examples

- The head of the European Commission, Ursula von der Leyen, has said the bloc is „ready to discuss" a US-backed proposal for a waiver on the patents for Covid-19 vaccines [...].
  theguardian.com
- Wie die Nachrichtenagentur Reuters unter Berufung auf die französische Journalistin Marie Carof-Gadel berichtet, fuhren französische Fischer am Donnerstagmorgen in den Hafen von Saint Helier auf der britischen Insel Jersey ein und positionierten ihr Boot dort unter anderem vor der Fähre „Commodore Goodwill". spiegel.de

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

Many options

- Rules/patterns
  - Over lemma and pos tags
  - Over dependency/phrase structure trees

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

Many options

- Rules/patterns
  - Over lemma and pos tags
  - Over dependency/phrase structure trees
- Classical machine learning (manual feature extraction code)
  - Hidden Markov Models
  - Conditional Random Fields

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

Many options

- Rules/patterns
  - Over lemma and pos tags
  - Over dependency/phrase structure trees
- Classical machine learning (manual feature extraction code)
  - Hidden Markov Models
  - Conditional Random Fields
- Classical deep learning (feed in embeddings)
  - Bilinear Long-Short-Term-Memory (Bi-LSTM)

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

Many options

- Rules/patterns
  - Over lemma and pos tags
  - Over dependency/phrase structure trees
- Classical machine learning (manual feature extraction code)
  - Hidden Markov Models
  - Conditional Random Fields
- Classical deep learning (feed in embeddings)
  - Bilinear Long-Short-Term-Memory (Bi-LSTM)
- Hot shit
  - Transformer-based model (BERT): Take trained BERT model, fine-tune to your task

How does NER work?

- Sequential problem: ‚Namedness' of one token depends on the previous one

→ It's a sequence labeling problem

  - Rules out: decision trees, SVMs, logistic regression, random forests, ...

Many options

- Rules/patterns
  - Over lemma and pos tags
  - Over dependency/phrase structure trees
- Classical machine learning (manual feature extraction code)
  - Hidden Markov Models
  - Conditional Random Fields
- Classical deep learning (feed in embeddings)
  - Bilinear Long-Short-Term-Memory (Bi-LSTM)
- Hot shit
  - Transformer-based model (BERT): Take trained BERT model, fine-tune to your task

In any case: You'll need annotated data – for training ML, and for testing everything

How does NER work?

Training

- You will not have enough training data to start from scratch

How does NER work?

Training

- You will not have enough training data to start from scratch
- Classical machine learning
  - Extract intelligent features using deep linguistic preprocessing

How does NER work?

Training

- You will not have enough training data to start from scratch
- Classical machine learning
  - Extract intelligent features using deep linguistic preprocessing
- Classical deep learning
  - Use pre-trained embeddings (GloVe/Word2Vec/FastText)
  - Pay attention to dates: Lot of new vocabulary
    - FastText embeddings work for OOV words, because they are (also) based on characters

How does NER work?

Training

- You will not have enough training data to start from scratch
- Classical machine learning
  - Extract intelligent features using deep linguistic preprocessing
- Classical deep learning
  - Use pre-trained embeddings (GloVe/Word2Vec/FastText)
  - Pay attention to dates: Lot of new vocabulary
    - FastText embeddings work for OOV words, because they are (also) based on characters
- Hot shit
  - Download pre-trained BERT model, fine-tune to your task
  - Unclear chances of success

NER Options

Decision Making

- How to make a decision between these options?

NER Options

Decision Making

- How to make a decision between these options?

Depends on your phenomenon!

- How lexicalised is it? How syntactically varied is it?
  - How many ways are there to express it in language?
  - Is it always expressed as a noun/noun phrase?
- How context-dependent is it?
  - Do the same tokens sometimes instantiate it and sometimes not?
- How easy is it to annotate?
  - Do you need to read a lot of text to recognise it?
  - How certain are you when you see an instance?

Getting Started

- Rule engine Apache UIMA Ruta
  - Run in Jupyter notebook: `https://pypi.org/project/sos-ruta/`

Getting Started

- Rule engine Apache UIMA Ruta
  - Run in Jupyter notebook: `https://pypi.org/project/sos-ruta/`
- Classical machine learning
  - Weka: `https://www.cs.waikato.ac.nz/ml/weka/`
    - Tutorial: `https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2`

Getting Started

- Rule engine Apache UIMA Ruta
  - Run in Jupyter notebook: `https://pypi.org/project/sos-ruta/`
- Classical machine learning
  - Weka: `https://www.cs.waikato.ac.nz/ml/weka/`
    - Tutorial: `https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2`
- Deep Learning: Keras `https://keras.io`
  - BiLSTM Tutorial:
    `https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/`
  - BERT fine-tuning for NER Tutorial:
    `https://skimai.com/how-to-fine-tune-bert-for-named-entity-recognition-ner/`

Questions?

References I

- Berendt, Bettina (2015). *An introduction to sentiment analysis and opinion mining*.
- Jurafsky, Dan/James H. Martin (2019). *Speech and Language Processing*. 3rd ed. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/.
- Manning, Christopher D./Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.