UNIVERSITÄT
ZU KÖLN

Computerlinguistische Experimente und Ziele

HS Anwendungen der Computerlinguistik

Nils Reiter
nils.reiter@uni-koeln.de

16. November 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Today

- Preoțiuc-Pietro u. a. (2019): „Automatically Identifying Complaints in Social Media“
- Panchendrarajan u. a. (2016): „Implicit Aspect Detection in Restaurant Reviews using Cooccurence of Words“

- Which one did you like better?
- Which one was easier to understand?

Today

- Preoțiuc-Pietro u. a. (2019): „Automatically Identifying Complaints in Social Media“
- Panchendrarajan u. a. (2016): „Implicit Aspect Detection in Restaurant Reviews using Cooccurence of Words“

- Which one did you like better?
- Which one was easier to understand?

- Very typical NLP papers
  - 8-9 pages, densely written
  - Structure: Abstract – Introduction – Related work – Data description/analysis – Experimental part – Conclusions – References

Today

- Preoţiuc-Pietro u. a. (2019): „Automatically Identifying Complaints in Social Media“
- Panchendrarajan u. a. (2016): „Implicit Aspect Detection in Restaurant Reviews using Cooccurence of Words“

- Which one did you like better?
- Which one was easier to understand?

- Very typical NLP papers
  - 8-9 pages, densely written
  - Structure: Abstract – Introduction – Related work – Data description/analysis – Experimental part – Conclusions – References
- My opinion: Preoţiuc-Pietro u. a. (2019) ‚better‘ than Panchendrarajan u. a. (2016)

Reading up on Details, Techniques, Methods

Abbreviation	Reference
Manning/Schütze, 1999	Christopher D. Manning/Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts und London, England: MIT Press
Jurafsky/Martin, 2023	Dan Jurafsky/James H. Martin (2023). *Speech and Language Processing*. 3. Aufl. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/

Tabelle: References to text books

Comprehension Questions

- 10-fold cross validation
- ROC AUC
- Maximum entropy classification
- Cohen's Kappa

The Tasks

Preoțiuc-Pietro u. a. (2019)

The Tasks

Preoţiuc-Pietro u. a. (2019)

- Target concept: Complaints
- Binary classification of tweets
- A tweet is positive, if it contains at least one complaint speech act
- No context dependency

The Tasks

Preoțiuc-Pietro u. a. (2019)

- Target concept: Complaints
- Binary classification of tweets
- A tweet is positive, if it contains at least one complaint speech act
- No context dependency

Panchendrarajan u. a. (2016)

The Tasks

Preoțiuc-Pietro u. a. (2019)

- Target concept: Complaints
- Binary classification of tweets
- A tweet is positive, if it contains at least one complaint speech act
- No context dependency

Panchendrarajan u. a. (2016)

- Target concept: Mentioned and reviewed aspects
- Multi-label classification of sentences
  - Not explicitly stated by the authors
- No context dependency

The Tasks

Preoţiuc-Pietro u. a. (2019)

- Target concept: Complaints
- Binary classification of tweets
- A tweet is positive, if it contains at least one complaint speech act
- No context dependency

Panchendrarajan u. a. (2016)

- Target concept: Mentioned and reviewed aspects
- Multi-label classification of sentences
  - Not explicitly stated by the authors
- No context dependency

Reminder: Classification

- Organize items into previously defined classes

Manning/Schütze, 1999, 192,575

- Multi-class: More than two classes (i.e., more than binary)
- Multi-label: Each item can get more than one label

The Data Sets

Preoțiuc-Pietro u. a. (2019)

The Data Sets

Preoţiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets

The Data Sets

Preoțiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets
- Preprocessing
  - Replace all usernames
  - Replace all URLs
  - Extract unigrams

The Data Sets

Preoțiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets
- Preprocessing
  - Replace all usernames
  - Replace all URLs
  - Extract unigrams
- Annotation
  - Two independent annotators
  - Agreement $\kappa = 0.731$ (Cohen, 1960)

The Data Sets

Preoțiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets
- Preprocessing
  - Replace all usernames
  - Replace all URLs
  - Extract unigrams
- Annotation
  - Two independent annotators
  - Agreement $\kappa = 0.731$ (Cohen, 1960)

Panchendrarajan u. a. (2016)

- 1000 restaurant reviews from Yelp

The Data Sets

Preoțiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets
- Preprocessing
  - Replace all usernames
  - Replace all URLs
  - Extract unigrams
- Annotation
  - Two independent annotators
  - Agreement $\kappa = 0.731$ (Cohen, 1960)

Panchendrarajan u. a. (2016)

- 1000 restaurant reviews from Yelp
- Annotation (p. 135)
  - Two independent annotators on 3 samples of 100 reviews
  - Sentence-wise annotation
  - Agreement $\kappa = 0.834$ (Cohen, 1960)

The Data Sets

Preoțiuc-Pietro u. a. (2019)

- 3449 English tweets, no retweets
  - 1971 to which support accounts replied
  - 739 @-replies
  - 739 other tweets
- Preprocessing
  - Replace all usernames
  - Replace all URLs
  - Extract unigrams
- Annotation
  - Two independent annotators
  - Agreement $\kappa = 0.731$ (Cohen, 1960)

Panchendrarajan u. a. (2016)

- 1000 restaurant reviews from Yelp
- Annotation (p. 135)
  - Two independent annotators on 3 samples of 100 reviews
  - Sentence-wise annotation
  - Agreement $\kappa = 0.834$ (Cohen, 1960)
- Highly skewed distribution (Most sentences do not contain implicit aspects)

Experimental Setup

Preoțiuc-Pietro u. a. (2019)

- 10-fold cross validation JM19, 69
- Parameters: 3-fold CV in inner loop

Experimental Setup

Preoțiuc-Pietro u. a. (2019)

- 10-fold cross validation JM19, 69
- Parameters: 3-fold CV in inner loop
- Evaluation
  - Mean accuracy
  - F1 (macro-average)
  - ROC AUC Manning/Schütze, 1999, 270
    - (ROC = receiver operating characteristic curve / AUC = area under curve)

Experimental Setup

Preoțiuc-Pietro u. a. (2019)

- 10-fold cross validation JM19, 69
- Parameters: 3-fold CV in inner loop
- Evaluation
  - Mean accuracy
  - F1 (macro-average)
  - ROC AUC Manning/Schütze, 1999, 270
    - (ROC = receiver operating characteristic curve / AUC = area under curve)

Panchendrarajan u. a. (2016)

- 10-fold cross validation JM19, 69
- Additional 400 reviews used for testing M1

Experimental Setup

Preoțiuc-Pietro u. a. (2019)

- 10-fold cross validation JM19, 69
- Parameters: 3-fold CV in inner loop
- Evaluation
  - Mean accuracy
  - F1 (macro-average)
  - ROC AUC Manning/Schütze, 1999, 270
    - (ROC = receiver operating characteristic curve / AUC = area under curve)

Panchendrarajan u. a. (2016)

- 10-fold cross validation JM19, 69
- Additional 400 reviews used for testing M1
- Evaluation
  - Precision/recall/F1 Manning/Schütze, 1999, 267 ff.

Preprocessing

Processing steps before actual task solving

Preoțiuc-Pietro u. a. (2019)

- Part of speech
- Sentiment
- Request detection
- Politeness
- Time expressions

Preprocessing

Processing steps before actual task solving

Preoțiuc-Pietro u. a. (2019)

- Part of speech
- Sentiment
- Request detection
- Politeness
- Time expressions
- Word2vec

Preprocessing

Processing steps before actual task solving

Preoțiuc-Pietro u. a. (2019)

- Part of speech
- Sentiment
- Request detection
- Politeness
- Time expressions
- Word2vec
- Rule-based ad-hoc systems
  - Intensifiers
  - Pronoun types
  - LIWC

Preprocessing

Processing steps before actual task solving

Preoțiuc-Pietro u. a. (2019)

- Part of speech
- Sentiment
- Request detection
- Politeness
- Time expressions
- Word2vec
- Rule-based ad-hoc systems
  - Intensifiers
  - Pronoun types
  - LIWC

Panchendrarajan u. a. (2016)

- Dependency relations
  - Which one?

Preprocessing

Processing steps before actual task solving

Preoțiuc-Pietro u. a. (2019)

- Part of speech
- Sentiment
- Request detection
- Politeness
- Time expressions
- Word2vec
- Rule-based ad-hoc systems
  - Intensifiers
  - Pronoun types
  - LIWC

Panchendrarajan u. a. (2016)

- Dependency relations
  - Which one?

Pre-Processing

- No global definition of what counts as pre-processing
- Context-dependent

Methods

Preoţiuc-Pietro u. a. (2019)

- Baseline: Most frequent class

Methods

Preoțiuc-Pietro u. a. (2019)

- Baseline: Most frequent class
- Logistic regression with manually specified features JM19, 75 ff.

Methods

Preoţiuc-Pietro u. a. (2019)

- Baseline: Most frequent class
- Logistic regression with manually specified features JM19, 75 ff.
- Neural networks with one-hot-encoded word vectors as input
  - MLP: Feedforward neural network JM19, 129 ff.
  - LSTM: Sequential classifier (word by word) JM19, 184 ff.

Methods

Preoțiuc-Pietro u. a. (2019)

- Baseline: Most frequent class
- Logistic regression with manually specified features JM19, 75 ff.
- Neural networks with one-hot-encoded word vectors as input
  - MLP: Feedforward neural network JM19, 129 ff.
  - LSTM: Sequential classifier (word by word) JM19, 184 ff.

Panchendrarajan u. a. (2016)

- M1 (for explicit aspects): Maximum entropy classifier with n-grams as features ($2 \leq n \leq 5$)
  $=$ Logistic regression JM19, 75 ff.

Methods

Preoțiuc-Pietro u. a. (2019)

- Baseline: Most frequent class
- Logistic regression with manually specified features JM19, 75 ff.
- Neural networks with one-hot-encoded word vectors as input
  - MLP: Feedforward neural network JM19, 129 ff.
  - LSTM: Sequential classifier (word by word) JM19, 184 ff.

Panchendrarajan u. a. (2016)

- M1 (for explicit aspects): Maximum entropy classifier with n-grams as features ($2 \leq n \leq 5$)
  $=$ Logistic regression JM19, 75 ff.
- M2 (for implicit aspects)
  - Training: Collect dictionary (called ‚model' by the authors)
  - Testing
    1. Generate candidates, based on score A_i (Eq. 1)
    2. Remove candidates according to rules (Fig. 1)
  - Modification 1 and 2 (p. 133)

Section 1

Preoțiuc-Pietro u. a. (2019)

Features

- Preoțiuc-Pietro u. a. (2019, Section 4)
- Concepts
  - Bag of words: Frequency of all words, irrespective of their ordering
  - TF*IDF: Way of weighting the words (instead of absolute counts) Manning/Schütze, 1999, 541 ff.
  - Word2Vec: Method to represent word meaning in high-dimensional vector space JM19, 110 ff.
  - Clustering: Each tweet is associated with a cluster, based on the word vectors
    - This generates 200 features!

Experiments

Three experiments

- Experiment 1
  - Variation in the feature sets and/or methods
  - Original data set used for train/test
- Experiment 2
  - Additional data generated through distant supervision
    - Idea: Use weakly correlated properties to induce annotations
  - Seven hashtags based on training data
  - $36\,436$ additional tweets (positive/negative)
    - Roughly ten times as many
  - Two ways to combine the data sets
- Experiment 3: Cross-domain

Experiment 1

Model	Acc	F1	AUC
Most frequent class	64.2	39.1	0.5
Sentiment – Stanford	68	55.6	0.696
Complaint Specific (all)	65.7	55.2	0.634
Downgraders	65.4	49.8	0.615
POS Bigrams	72.2	66.8	0.756
LIWC	71.6	65.8	0.784
Word2Vec Clusters	67.7	58.3	0.738
Bag-of-Words	79.8	77.5	0.866
All	80.5	78	0.873
MLP	78.3	76.2	0.845
LSTM	80.2	77	0.864

Tabelle: Experimental Results (excerpt)

Experiment 2

Model	Acc	F1	AUC
Most Frequent Class	64.2	39.1	0.5
LR-All Features – Original Data	80.5	78	0.873
Dist. Supervision + Pooling	77.2	75.7	0.853
Dist. Supervision + EasyAdapt	81.2	79	0.885

Tabelle: Results of Experiment 2

Conclusions

- Concept rooted in linguistic work
- Created data set
- Analysis of data set
- Predictive model with reasonable performance

Section 2

Panchendrarajan u. a. (2016)

Experiments

Two experiments

- Experiment 1: Comparison of prediction performance of different settings
- Experiment 2: Isolated evaluation on sentences with two/more than two aspects

Experimental Results

M1	M2	Precision	Recall	F1
Oracle	As described	0.947	0.758	0.842
	Schouten et al.	0.495	0.929	0.645
	Modification 1	0.916	0.752	0.826
	Modification 2	0.931	0.754	0.834
M1	As described	0.886	0.694	0.779

Tabelle: Experimental Results. Oracle: Assume that a vital preprocessing step works perfectly.

It can be seen in Table 1 that our approach gives the best result. Moreover it is worth noting that the precision drops drastically from 0.947 to 0.529 in Modification 2 as it does not execute Step 2. (Panchendrarajan u. a., 2016, 135)

Experimental Results

M1	M2	Precision	Recall	F1
Oracle	As described	0.947	0.758	0.842
	Schouten et al.	0.495	0.929	0.645
	Modification 1	0.916	0.752	0.826
	Modification 2	0.931	0.754	0.834
M1	As described	0.886	0.694	0.779

Tabelle: Experimental Results. Oracle: Assume that a vital preprocessing step works perfectly.

It can be seen in Table 1 that our approach gives the best result. Moreover it is worth noting that the precision drops drastically from 0.947 to 0.529 in Modification 2 as it does not execute Step 2. (Panchendrarajan u. a., 2016, 135)

Experimental Results

M1	M2	Precision	Recall	F1
Oracle	As described	0.947	0.758	0.842
	Schouten et al.	0.495	0.929	0.645
	Modification 1	0.916	0.752	0.826
	Modification 2	0.931	0.754	0.834
M1	As described	0.886	0.694	0.779

Tabelle: Experimental Results. Oracle: Assume that a vital preprocessing step works perfectly.

It can be seen in Table 1 that our approach gives the best result. Moreover it is worth noting that the precision drops drastically from 0.947 to 0.529 in Modification 2 as it does not execute Step 2. (Panchendrarajan u. a., 2016, 135)

- Automatically achieved results only in last row

Experimental Results

M1	M2	Precision	Recall	F1
Oracle	As described	0.947	0.758	0.842
	Schouten et al.	0.495	0.929	0.645
	Modification 1	0.916	0.752	0.826
	Modification 2	0.931	0.754	0.834
M1	As described	0.886	0.694	0.779

Tabelle: Experimental Results. Oracle: Assume that a vital preprocessing step works perfectly.

It can be seen in Table 1 that our approach gives the best result. Moreover it is worth noting that the precision drops drastically from 0.947 to 0.529 in Modification 2 as it does not execute Step 2. (Panchendrarajan u. a., 2016, 135)

- Mismatch between text and table

Summary

- Typical NLP papers: Focus in methods
- Complaints
  - Very clear
  - Classical machine learning wins
- Reviews
  - Implicit aspects in restaurant reviews
  - Machine learning and rules on top