Assessing Information Quality Facets in Blogs and Web Pages

Elisabeth Lex

Know Center Graz

15.09.2010

http://www.**know-center**.at

Agenda

- Introduction
- Blog Classification: Information Quality Facets
- Features, Example Facets
- ECML challenge
- Lessons Learned

Introduction

- On the Web: huge amount of content

  → Navigating this content not an easy task

- Transformation especially in the media domain
  - Breaking news often in blogosphere or over twitter before published by traditional media agencies
  - Much more up to date
  - Wide audience, open to (almost) anyone
  - Powerful e.g. for media resonance analysis

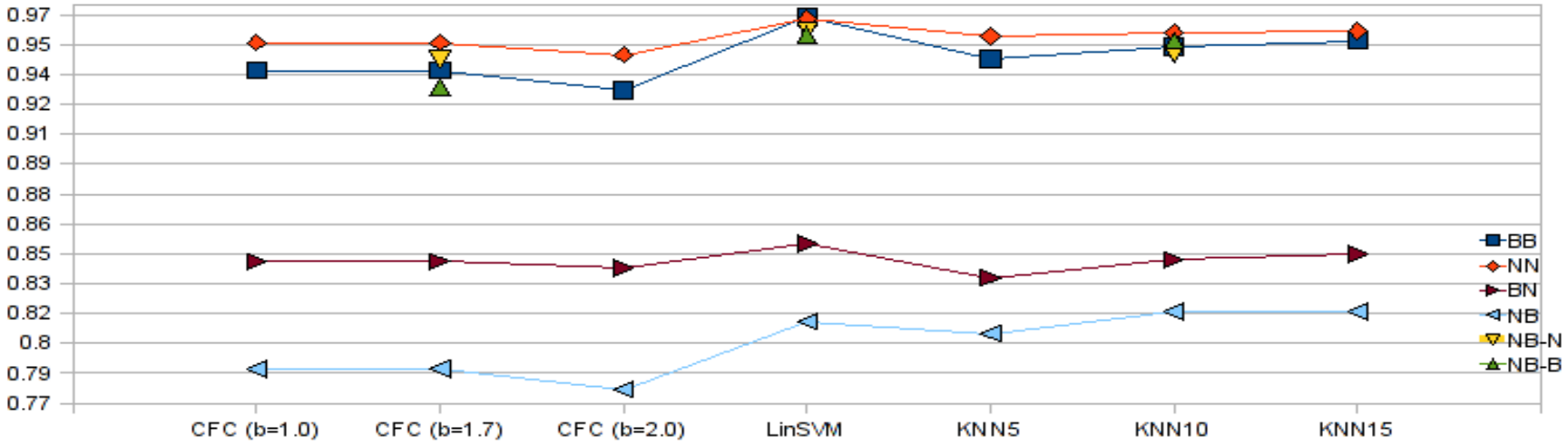
Organize news related blogs...

- Use blog tagging information?
  - However, tagging vocabulary not consistent
  - Subjective, not commonly agreed upon
- Use News Categories (Colon-Classification (CC))
  - Assigned by experts
  - Common categories like Politics, Sports

→ Exploit this high quality labels to categorize blogs!

Blog Organization: Cross Domain Classification Approach

- 5 categories: Politics, Economy, Sports, Culture, Science
- News Corpus: 5600 examples from German articles
- Blog Corpus: 2800 Politic blog posts, 2800 economy, 2400 sports, 1800 culture, 1100 science



Lex, E., Seifert, C., Granitzer, M., and Juffinger, A., 2009. Automated Blog Classification: A Cross Domain Approach. In *Proceedings of IADIS International Conference WWW/Internet.*

Example: US elections 2008

Sarah Palin's Expensive Clothes

James Joyner | *Wednesday, October 22, 2008*

We've had John Edwards' haircuts, John McCain's shoes, Michelle Obama's snacks (a story that turned out to be untrue), and now, Sarah Palin's wardrobe.

> The Republican National Committee has spent more than $150,000 to clothe and accessorize vice presidential candidate Sarah Palin and her family since her surprise pick by John McCain in late August.
>
> According to financial disclosure records, the accessorizing began in early September and included bills from Saks Fifth Avenue in St. Louis and New York for a combined $49,425.74. The records also document a couple of big-time shopping trips to Neiman Marcus in Minneapolis, including one $75,062.63 spree in early September.
>
> The RNC also spent $4,716.49 on hair and makeup through September after reporting no such costs in August.

Sarah Palin, in a red leather jacket, waves as she steps on stage before a crowd at a baseball field in Grand Junction, Colo., on Monday.

0 tweets

tweet

Jim Treacher thinks all these stories are a distraction from the real issues. And, of course, they are. But these are the sort of stories that seem to resonate with voters.

Matt Yglesias wonders why these expenditures are legal, noting "this seems to open the door to candidates using party committee money as a personal slush fund." Kevin Drum snarks that, "I'm sure that after the campaign is over the RNC plans to donate the clothing to homeless shelters in small towns around the country where they don't have stores like Saks or Barney's." Really, this is no worse than Al Gore spending tens of thousands having Naomi Wolfe tell him to wear "earth tones."

Why, precisely, it costs that kind of money to outfit Sarah Palin in a different red outfit every day, I haven't a clue. Presumably, it's more expensive to dress a woman — McCain and Obama are pretty much required to wear only dark suits and solid white or blue shirts — but this is expensive. Not to mention the fact that down-to-earth hockey moms — and pit bulls, for that matter — don't shop at Saks and Neiman Marcus. Then again, they're not typically plucked out of virtual obscurity and thrust into the national spotlight right before a national political convention, either.

The Obama Plan: Economists Speak Out on Forbes

Forbes brings us an interesting piece by a pair of research associates at the National Bureau of Economic Research. The topic, unsurprisingly, is health care and the approach taken on the subject by the Democratic nominee for U.S. president.

Example: US elections 2008

Sarah Palin's Expensive Clothes

James Joyner | *Wednesday, October 22, 2008*

We've had John Edwards' haircuts, John McCain's shoes, Michelle Obama's snacks (a story that turned out to be untrue), and now, Sarah Palin's wardrobe.

> The Republican National Committee has spent more than $150,000 to clothe and accessorize vice presidential candidate Sarah Palin and her family since her surprise pick by John McCain in late August.
>
> According to financial disclosure records, the accessorizing began in early September and included bills from Saks Fifth Avenue in St. Louis and New York for a combined $49,425.74. The records also document a couple of big-time shopping trips to Neiman Marcus in Minneapolis, including one $75,062.63 spree in early September.
>
> The RNC also spent $4,716.49 on hair and makeup through September after reporting no such costs in August.

Sarah Palin, in a red leather jacket, waves as she steps on stage before a crowd at a baseball field in Grand Junction, Colo., on Monday.

0 tweets tweet

Jim Treacher thinks all these stories are a distraction from the real issues. And, of course, they are. But these are the sort of stories that seem to resonate with voters.

Matt Yglesias wonders why these expenditures are legal, noting "this seems to open the door to candidates using party committee money as a personal slush fund." Kevin Drum snarks that, "I'm sure that after the campaign is over the RNC plans to donate the clothing to homeless shelters in small towns around the country where they don't have stores like Saks or Barney's." Really, this is no worse than Al Gore spending tens of thousands having Naomi Wolfe tell him to wear "earth tones."

Why, precisely, it costs that kind of money to outfit Sarah Palin in a different red outfit every day, I haven't a clue. Presumably, it's more expensive to dress a woman — McCain and Obama are pretty much required to wear only dark suits and solid white or blue shirts — but this is expensive. Not to mention the fact that down-to-earth hockey moms — and pit bulls, for that matter — don't shop at Saks and Neiman Marcus. Then again, they're not typically plucked out of virtual obscurity and thrust into the national spotlight before a national political convention, either.

Sarah Palin's clothes

The Obama Plan: Economists Speak Out on Forbes

Forbes brings us an interesting piece by a pair of research associates at the National Bureau of Economic Research. The topic, unsurprisingly, is health care and the approach taken on the subject by the Democratic nominee for U.S. president.

Economy and Health care

Example: US elections 2008

Sarah Palin's Expensive Clothes

James Joyner | *Wednesday, October 22, 2008*

We've had John Edwards' haircuts, John McCain's shoes, Michelle Obama's snacks (a story that turned out to be untrue), and now, Sarah Palin's wardrobe.

> The Republican National Committee has spent more than $150,000 to clothe and accessorize vice presidential candidate Sarah Palin and her family since her surprise pick by John McCain in late August.
>
> According to financial disclosure records, the accessorizing began in early September and included bills from Saks Fifth Avenue in St. Louis and New York for a combined $49,425.74. The records also document a couple of big-time shopping trips to Neiman Marcus in Minneapolis, including one $75,062.63 spree in early September.
>
> The RNC also spent $4,716.49 on hair and makeup through September after reporting no such costs in August.

Sarah Palin, in a red leather jacket, waves as she steps on stage before a crowd at a baseball field in Grand Junction, Colo., on Monday.

0 tweets tweet

Jim Treacher thinks all these stories are a distraction from the real issues. And, of course, they are. But these are the sort of stories that seem to resonate with voters.

Matt Yglesias wonders why these expenditures are legal, noting "this seems to open the door to candidates using party committee money as a personal slush fund." Kevin Drum snarks that, "I'm sure that after the campaign is over the RNC plans to donate the clothing to homeless shelters in small towns around the country where they don't have stores like Saks or Barney's." Really, this is no worse than Al Gore spending tens of thousands having Naomi Wolfe tell him to wear "earth tones."

Why, precisely, it costs that kind of money to outfit Sarah Palin in a different red outfit every day, I haven't a clue. Presumably, it's more expensive to dress a woman — McCain and Obama are pretty much required to wear only dark suits and solid white or blue shirts — but this is expensive. Not to mention the fact that down-to-earth hockey moms — and pit bulls, for that matter — don't shop at Saks and Neiman Marcus. Then again, they're not typically plucked out of virtual obscurity and thrust into the national spotlight before a national political convention, either.

The Obama Plan: Economists Speak Out on Forbes

Forbes brings us an interesting piece by a pair of research associates at the National Bureau of Economic Research. The topic, unsurprisingly, is health care and the approach taken on the subject by the Democratic nominee for U.S. president.

Blogs provides different aspects!

Sarah Palin's clothes

Economy and Health care

Ergo…

- Blogs provide different facets to a topic
- Rich source for media resonance analysis, reputation tracking, blog distillation (TREC)

→ Extract facets – which features?

→ Non topic oriented facets (emotionality)

  - → Useful for addressing personal information needs
  - → Can be exploited for quality estimation
  - → Useful for Web archival, triggering of crawling processes

Ergo…

- Blogs provide different facets to a topic
- Rich source for media resonance analysis, reputation tracking, blog distillation (TREC)

→ Extract facets – which features?

→ Non topic oriented facets (emotionality)

  - → Useful for addressing personal information needs
  - → Can be exploited for quality estimation
  - → Useful for Web archival, triggering of crawling processes

BUT: which features are suited to get the non topic facets?

Lexical Features vs Stylometric Features

- Common Bag of Words features:
  - Unigrams, Bigrams, Trigrams
  - Stems
  - Nouns, Verbs, Adjectives
  - Leading and Trailing Graphems
  - Personal Pronouns

- Stylometric features
  - Punctuation, Emoticons
  - Words in sentences
  - Chars in sentences
  - Noun+verb sentences (complete sentences)
  - Avg number of unique pos tags
  - Lower case/upper case
  - Word length
  - Adjective rate and adverb rate

Lexical Features vs Stylometric Features

- Common Bag of Words features:
  - Unigrams, Bigrams, Trigrams
  - Stems
  - Nouns, Verbs, Adjectives
  - Leading and Trailing Graphems
  - Personal Pronouns

- Stylometric features
  - Punctuation, Emoticons
  - Words in sentences
  - Chars in sentences
  - Noun+verb sentences (complete sentences)
  - Avg number of unique pos tags
  - Lower case/upper case
  - Word length
  - Adjective rate and adverb rate

TOPIC INDEPENDENT

Experiment: Objectivity of Online News (1/3)

- Corpus: British Newspaper:
  - 2 high quality (The Telegraph, The Guardian, 3000 articles)
  - 2 yellow press (The Sun, The Daily Mail, 2500 articles)
  - 6 categories: Columnist, Royals, The Royal Family, Diana, Music, Film, Celebrity News, and Bollywood

Facet Objectivity (2/3)

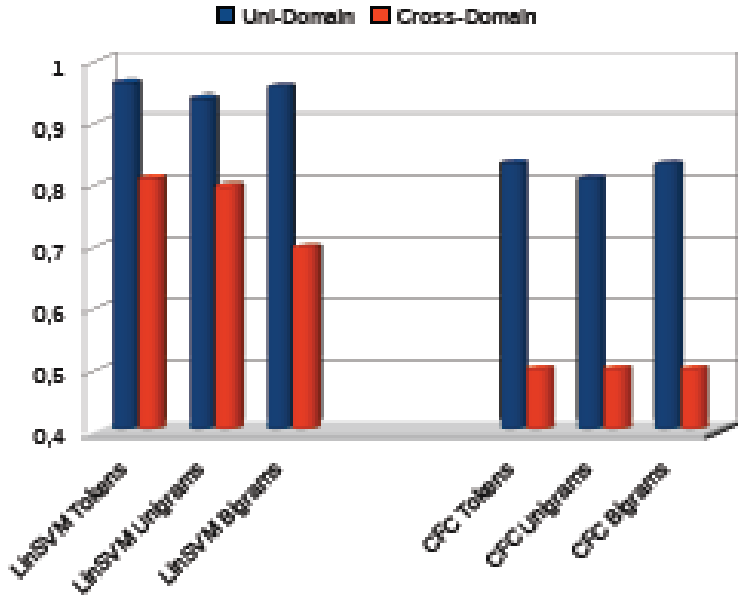


Figure 1: Performance on News: Word based Classifier for Cross-Domain/Category

Facet Objectivity (3/3)

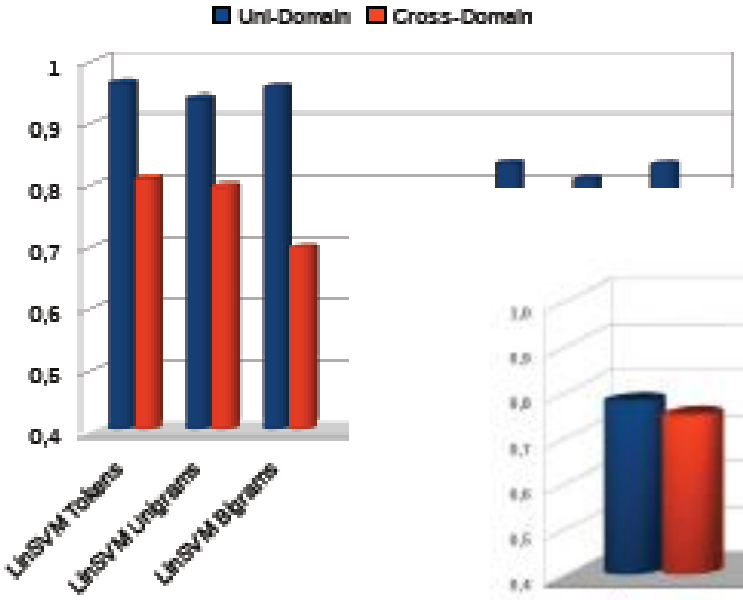


Figure 1: Performance o
sifier for Cross-Domain/

Figure 3: Performance on News: Style Classifier for Cross-Domain/Category

Lex, E., Juffinger, A., and Granitzer, M. Objectivity Classification in Online Media. HT 2010

Experiments: Emotion Classification - Dataset

- Manually annotated subset of Blogs08 TREC dataset: 83 annotated blogs (12844 Blog entries)

	News Related	Other
blog level	29%	71%
entry level	30%	70%

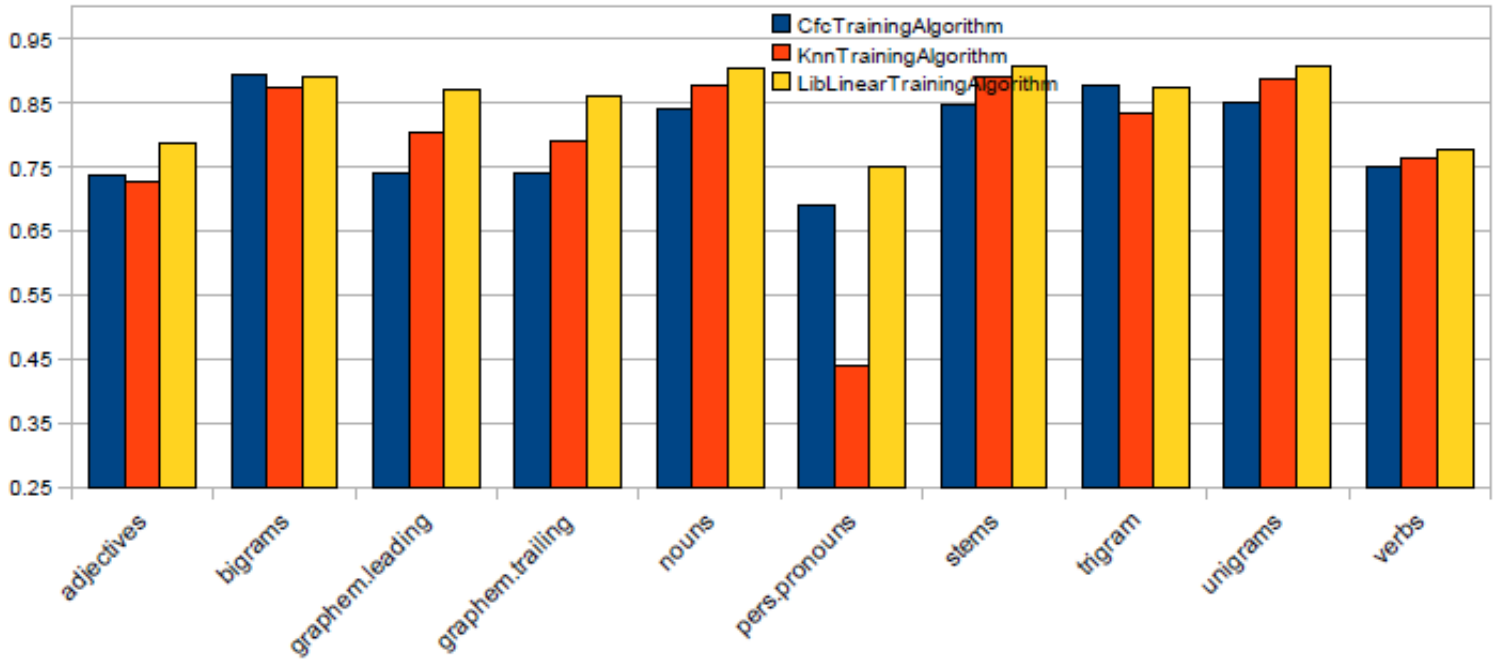
(a) News vs. Rest

	Emotional	Neutral
blog level	52%	48%
entry level	40%	60%

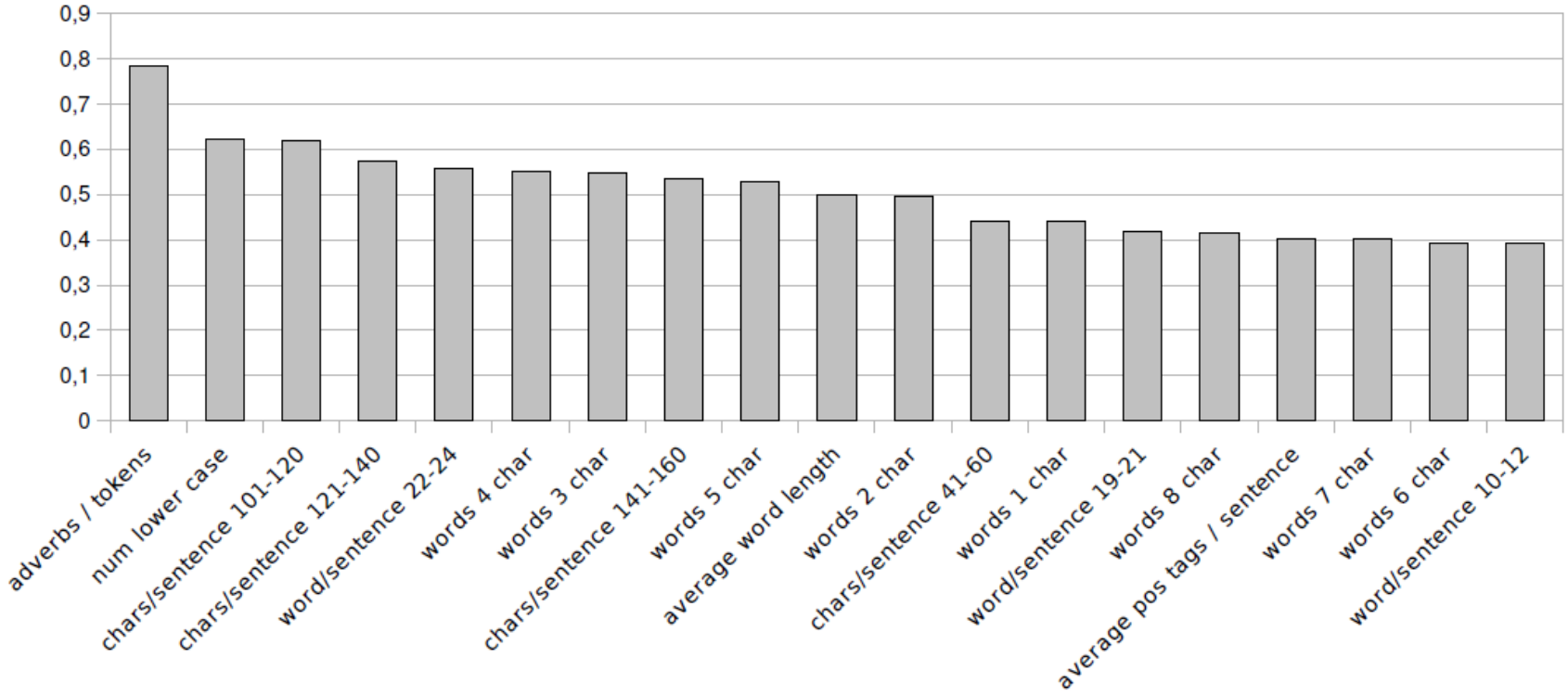
(b) Emotion Classification Task

Table I
CORPUS DISTRIBUTIONS

Facet Emotionality – Lexical Features

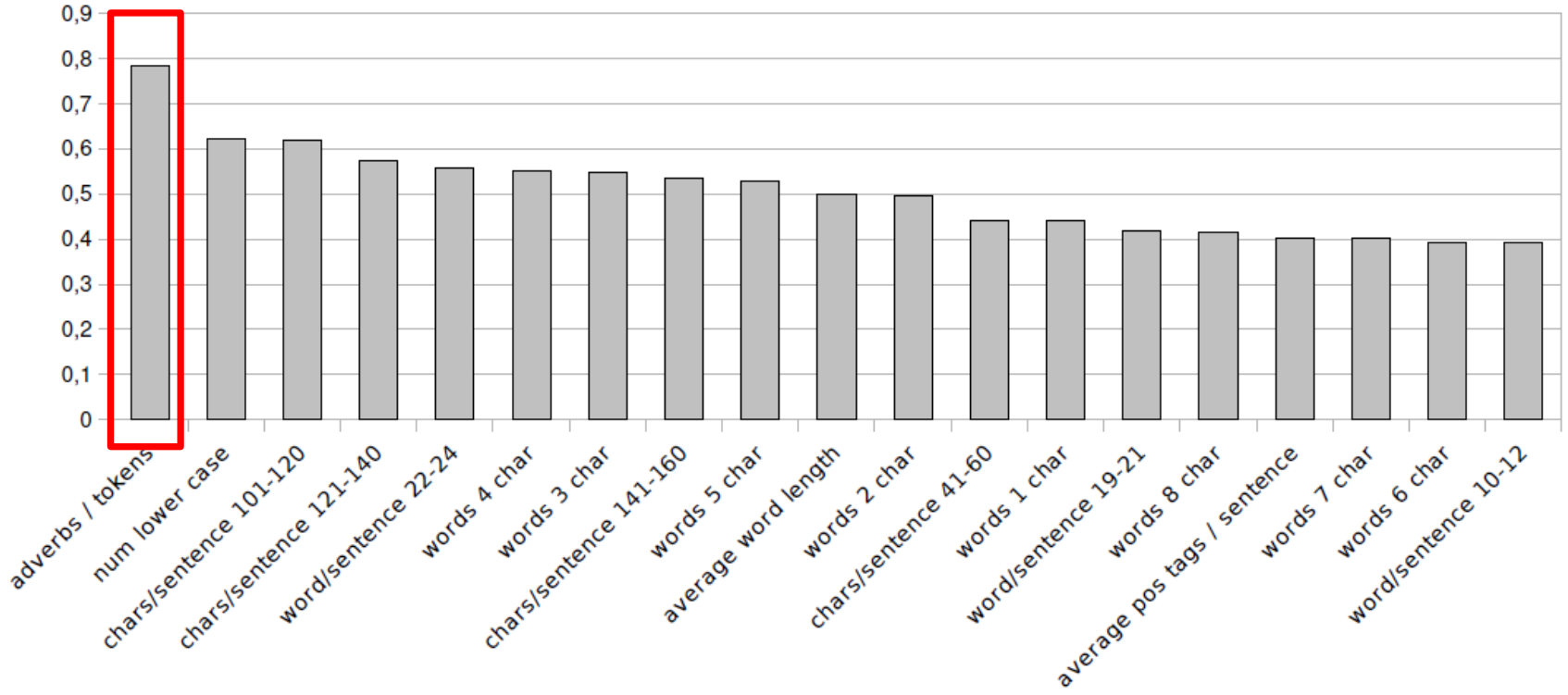


Mutual Information for Stylometric Features



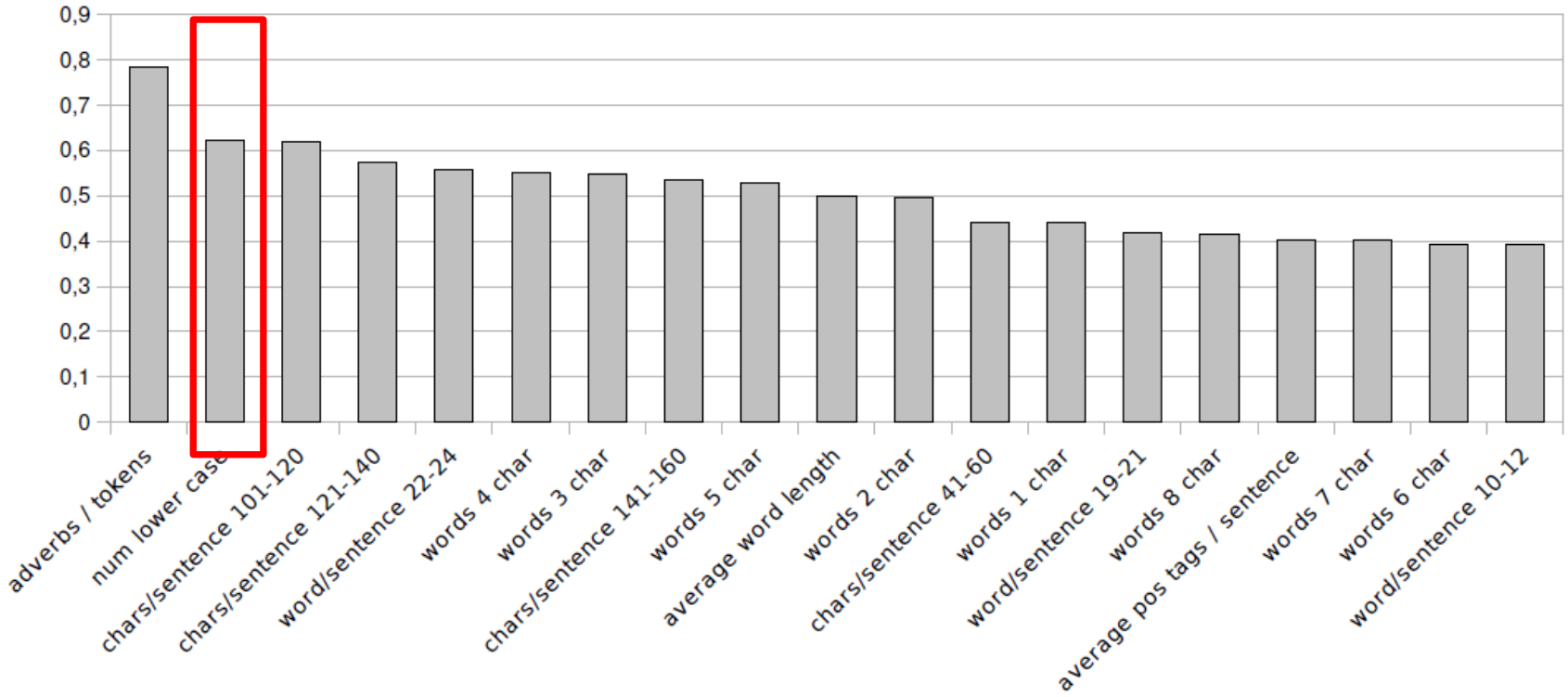
(b) Emotion Classification Task

Mutual Information for Stylometric Features



(b) Emotion Classification Task

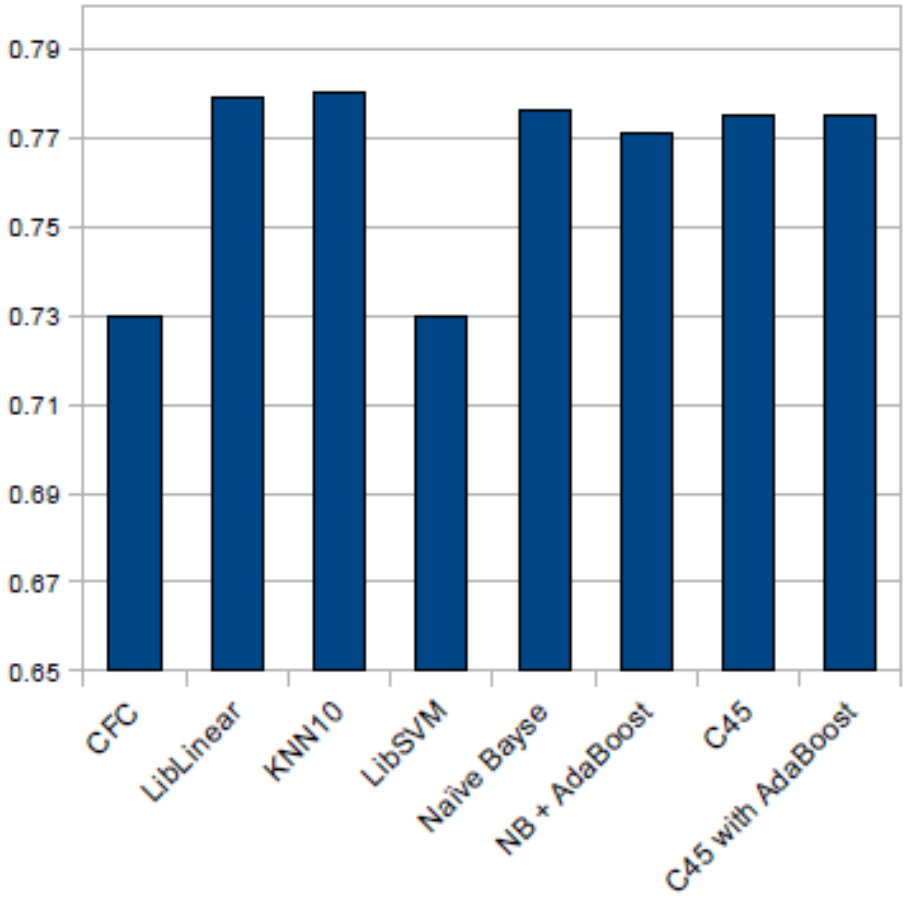
Mutual Information for Stylometric Features



(b) Emotion Classification Task

Facet Emotionality – Stylometric Features

- Lower accuracy
- BUT: Topic independent!



Lex, E., Juffinger, A., and Granitzer, M. A Comparison of Stylometric and Lexical Features for Emotion Classification in Blogs. TIR 2010

ECML Discovery Challenge

- 3 Tasks: Web Genre, Information Quality Facets, Quality
  - Multilingual quality task: aggregate function of genre and facets
    - Web Spam, News/Editorial, Commercial, Educational/Research, Discussion, Personal/Leisure
    - Trustworthiness, Bias, Neutrality (intrinsic content quality)

    → Results: Web hosts ranked by NDCG

Utility Score

```
utilityScore = 0;
if (News-Edit OR Educational) {
value = 5;
} else if (Discussion) {
value = 4;
} else if (Commercial OR Personal-Leisure) {
value = 3;
}
if (neutrality == 3) value += 2;
if (bias == 1) value -= 2;
if (trustworthiness == 3) value +=2;
```

Utility Score

```
utilityScore = 0;
if (News-Edit OR Educational) {
value = 5;
} else if (Discussion) {
value = 4;
} else if (Commercial OR Personal-Leisure) {
value = 3;
}
if (neutrality == 3) value += 2;
if (bias == 1) value -= 2;
if (trustworthiness == 3) value +=2;
```

Use Case: Web Archival

ECML Discovery Challenge - Setup

- Features:
  - Link Features (176)
  - Content Features (95)
  - NLP Features (180)
  - Term frequencies (50 000 terms)

23M pages
99000 hosts

Table 1: Number of training samples

Category	Positive Samples [%]	Negative Samples [%]
WebSpam	4	96
News/Editorial	4.7	95.3
Educational/Research	43	57
Personal/Leisure	23.7	76.3
Commercial	45.4	54.6
Discussion	5.3	94.7
Bias	1.7	98.3
Neutrality	96.6	3.4
Trustworthiness	98.1	1.9

ECML Discovery Challenge

- Features:
  - Link Features
  - Content Features
  - NLP Features
  - Term frequencies

→Ensemble classifier approach
→SVM, Decision Tree (J48 with SMOTE), CFC

Table 1: Number of training samples

Category	Positive Samples [%]	Negative Samples [%]
WebSpam	4	96
News/Editorial	4.7	95.3
Educational/Research	43	57
Personal/Leisure	23.7	76.3
Commercial	45.4	54.6
Discussion	5.3	94.7
Bias	1.7	98.3
Neutrality	96.6	3.4
Trustworthiness	98.1	1.9

ECML Discovery Challenge: Results

Table 2: Results for Task 1

Category	NDCG
WebSpam	0.473
News/Editorial	0.416
Commercial	0.694
Educational/Research	0.688
Discussion	0.531
Personal/Leisure	0.583
Trustiness	0.397
Bias	0.540
Neutrality	0.51
Average	0.537

ECML Discovery Challenge: Results

Table 2: Results for Task 1

Category	NDCG
WebSpam	0.473
News/Editorial	0.416
Commercial	0.694
Educational/Research	0.688
Discussion	0.531
Personal/Leisure	0.583
Trustiness	0.397
Bias	0.540
Neutrality	0.51
Average	0.537

ECML Discovery Challenge: Results

Table 2: Results for Task 1

Category	NDCG
WebSpam	0.473
News/Editorial	0.416
Commercial	0.694
Educational/Research	0.688
Discussion	0.531
Personal/Leisure	0.583
Trustiness	0.397
Bias	0.540
Neutrality	0.51
Average	0.537

Table 3: Results for Task 2

Language	NDCG
English	0.844

ECML Discovery Challenge: Results

Table 2: Results for Task 1

Category	NDCG
WebSpam	0.473
News/Editorial	0.416
Commercial	0.694
Educational/Research	0.688
Discussion	0.531
Personal/Leisure	0.583
Trustiness	0.397
Bias	0.540
Neutrality	0.51
Average	0.537

Table 3: Results for Task 2

Language	NDCG
English	0.844

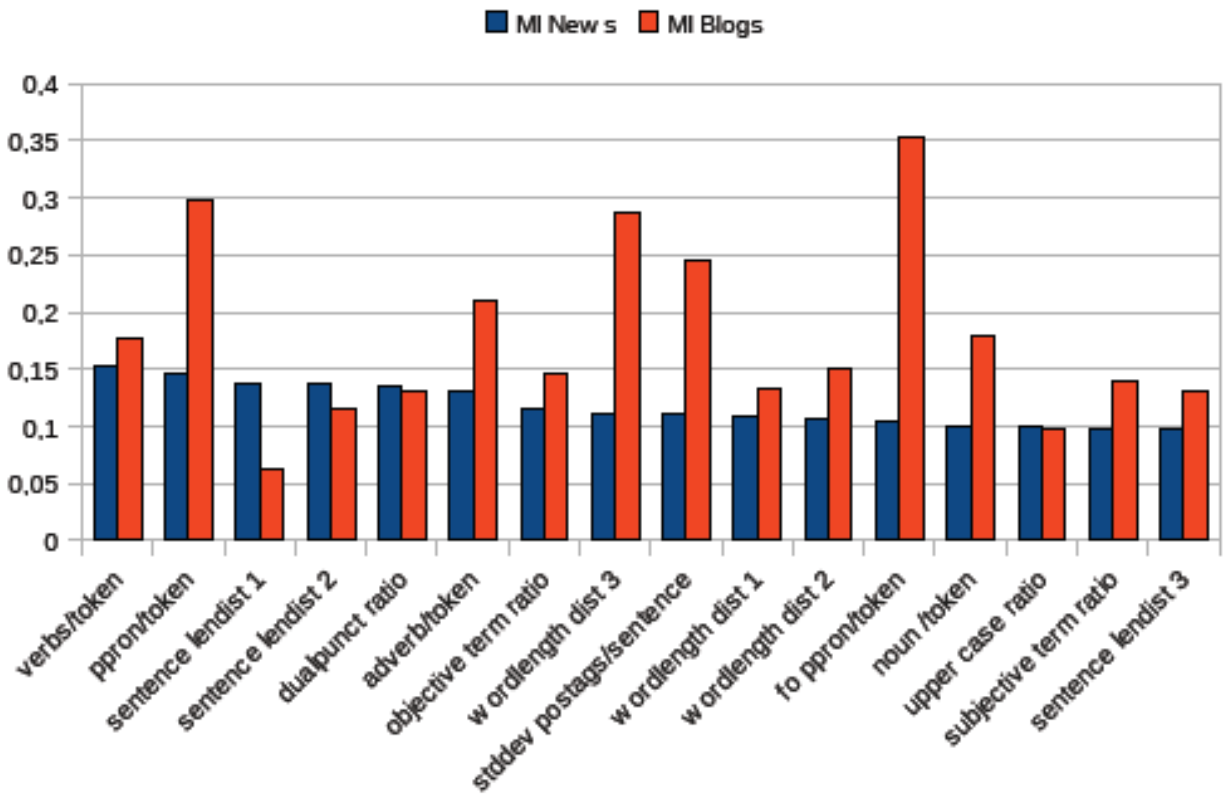
Table 4: Results for Task 3

Language	NDCG
German	0.792
French	0.823

Conclusions

- Information facets can be assessed from content with stylometric features
- Genre classification still a hard problem
  - Often, topics are learned
- Unbalanced datasets are very challenging
- Information quality is made up of multiple dimensions → depends on application context

Thank you for your attention!
Questions?


MI News
MI Blogs
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
verbs/token
ppron/token
sentence lendist 1
sentence lendist 2
dualpunct ratio
adverb/token
objective term ratio
w ordlength dist 3
stddev postags/sentence
w ordlength dist 1
w ordlength dist 2
fo ppron/token
noun /token
upper case ratio
subjective term ratio
sentence lendist 3

Blog Distillation Task: TREC challenge 2009

- Task performed on Blogs08 collection
- Task addressed quality aspects of retrieved blogs
- Mimicked an exploratory search task
  - E.g.: "Find me a *good* blog with a principal, recurring interest in X
- Facets were given – classify blogs into facets and rank retrieved blogs by relevance
  - Opinionated vs. factual
  - Personal vs. official
  - In-depth vs. shallow

TREC 2009 Blog Distillation Task Results

- Manually annotated 83 blogs into facets
- Indexed only 604k blogs out of 1.3 Mio.
- Submitted 3 runs on 3 different features: Nouns, Punctuation, Sentence statistics

→ Punctuation and sentence statistics are topic-independent!

→Best run with punctuation features!

  - → 5th out of 9 groups for facet "Personal"
  - → 6th out of 9 groups for facet "Opinionated"

Lex, E., Granitzer, M., and Juffinger, A., 2009. Facet Classication of Blogs: Know-Center at the TREC 2009 Blog Distillation Task