Unsupervised Detection of Anomalous Text

Part 2

Jozef Štyrák

Contents

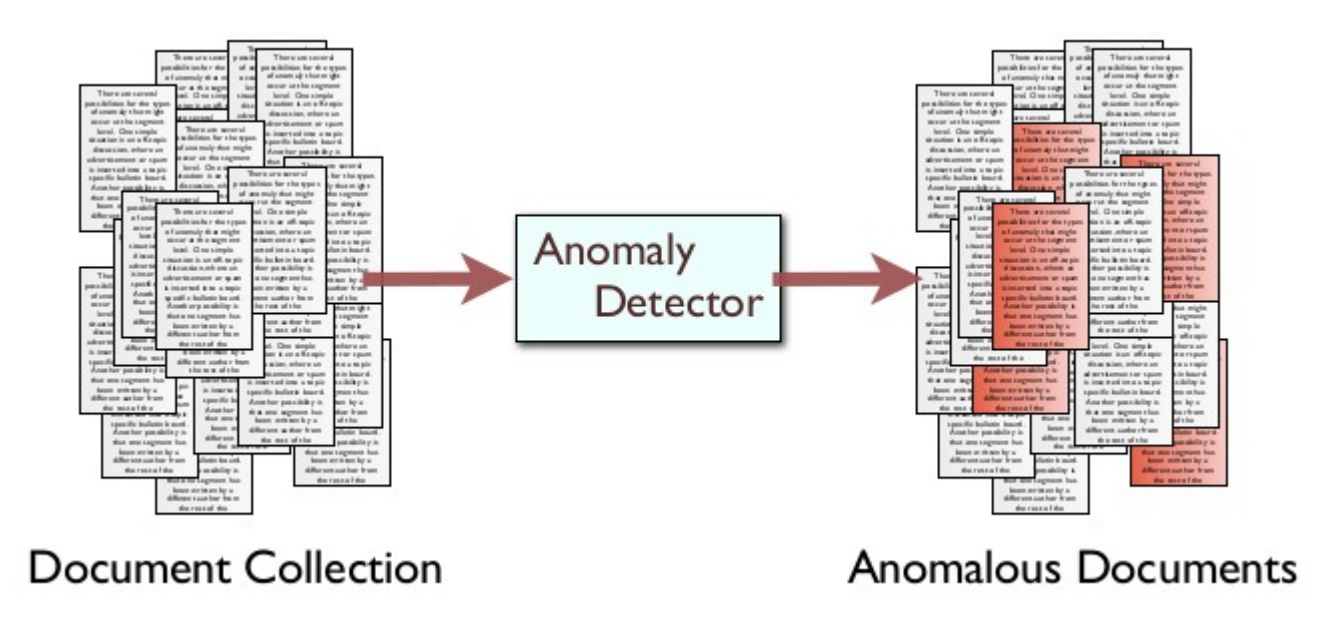
- Quick Recap
- Experimental Results
- IS and MUNI Applications

About

- Written by David Guthrie
- Doctor Degree Thesis at University of Sheffield in 2008

What is an Anomaly?

- “something that deviates from what is standard, normal or expected”
- Anomalies in text on segment level



Problem Definition

- Unsupervised Anomaly Detection
- Text is represented by set of features
- Text is split into segments, for each is counted a score – how much it deviates from what is normal for a particular document

Used Techniques

- ClustDist
- SDEDist
- Pcout
- MeanComp
- **TxtCompDist –** textual complement
- Baseline – choosing randomly

TxtCompDist Algorithm

- Measures distance from the textual complement
  - *X* – feature vector for the segment
  - *C* – feature vector for segment's complement

$$TxtCompDist(x, \boldsymbol{V}) = d(x, c_x)$$

- Novel method designed by authors
- In comparision with MeanComp better usage of ranked lists features (POS trigrams, adverbs,...)

Stahel-Donoho Estimator Distance

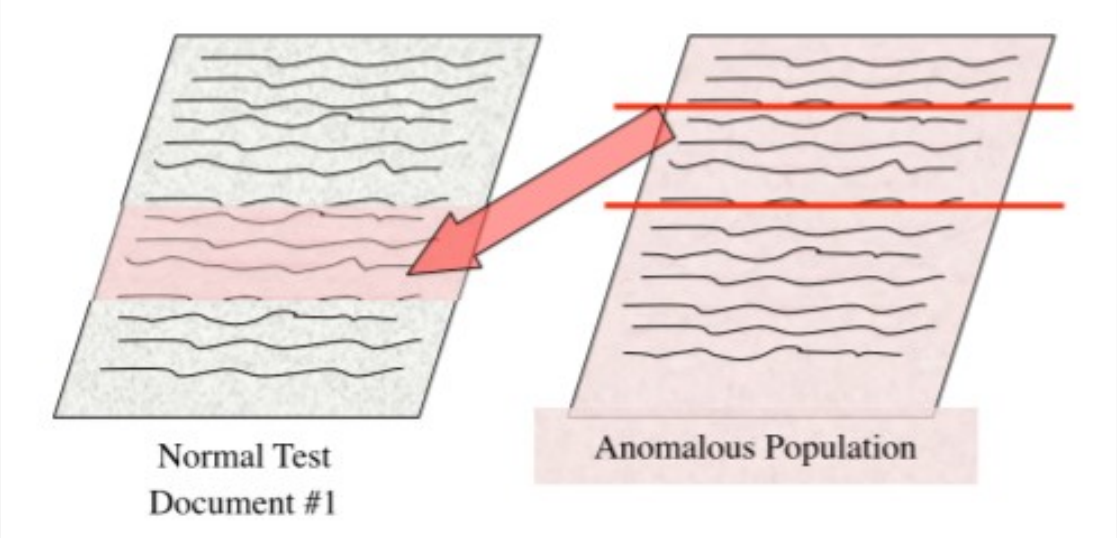
- Idea: to find projections of the data which maximize an observations distance from the center of the observations

$$SDEDist\left(x, \boldsymbol{V}\right) = max_a \frac{x^T a - median\left(\boldsymbol{V}\, a\right)}{mad\left(\boldsymbol{V}\, a\right)}$$

- Problem to find set of vectors *a*

Experimental Setup

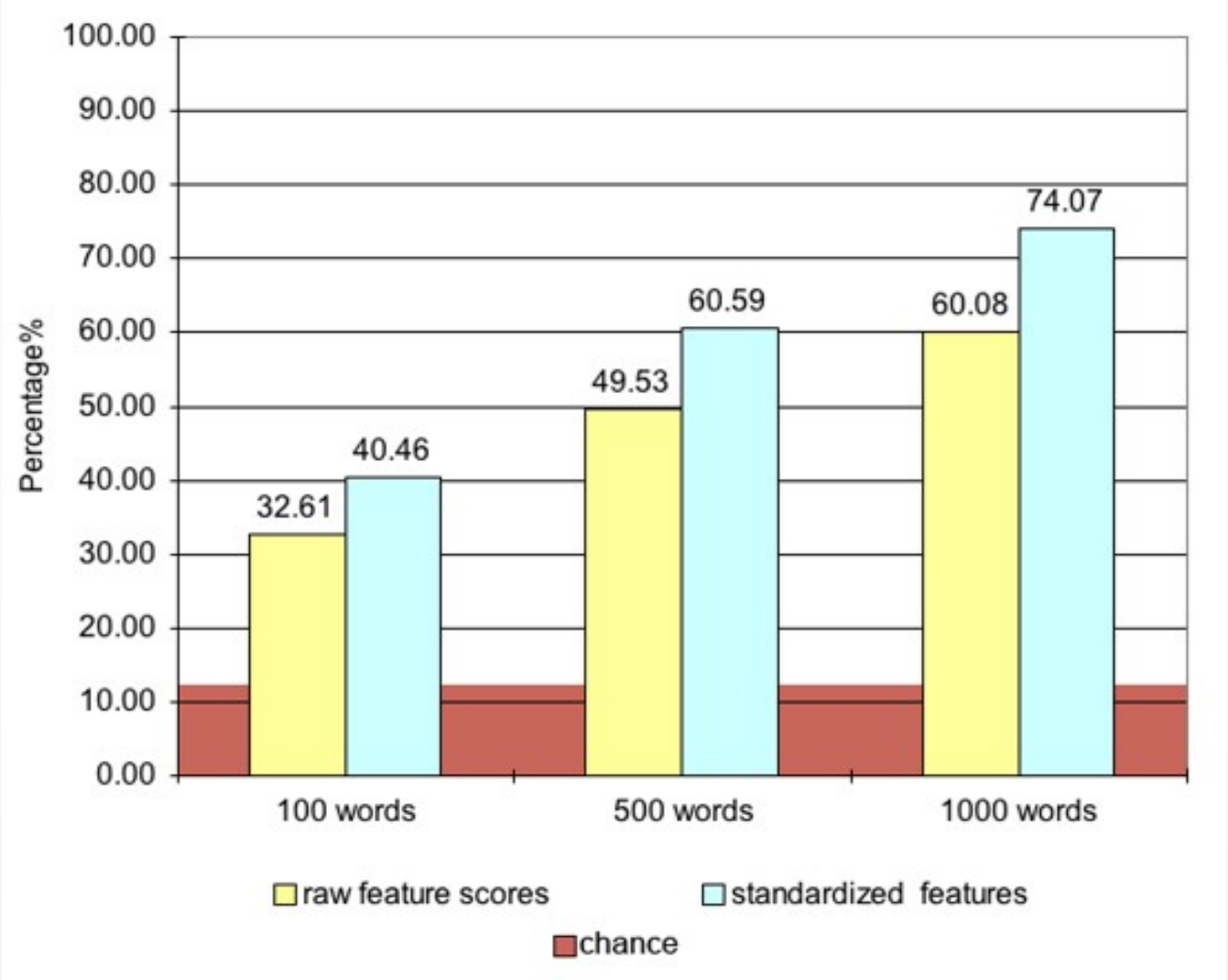
- Artificially created Test Document
  - 50 normal segments
  - 1 anomalous segment
- Output: List of segments ranked by how anomalous they are with respect to whole Test Document



Types of Anomalies

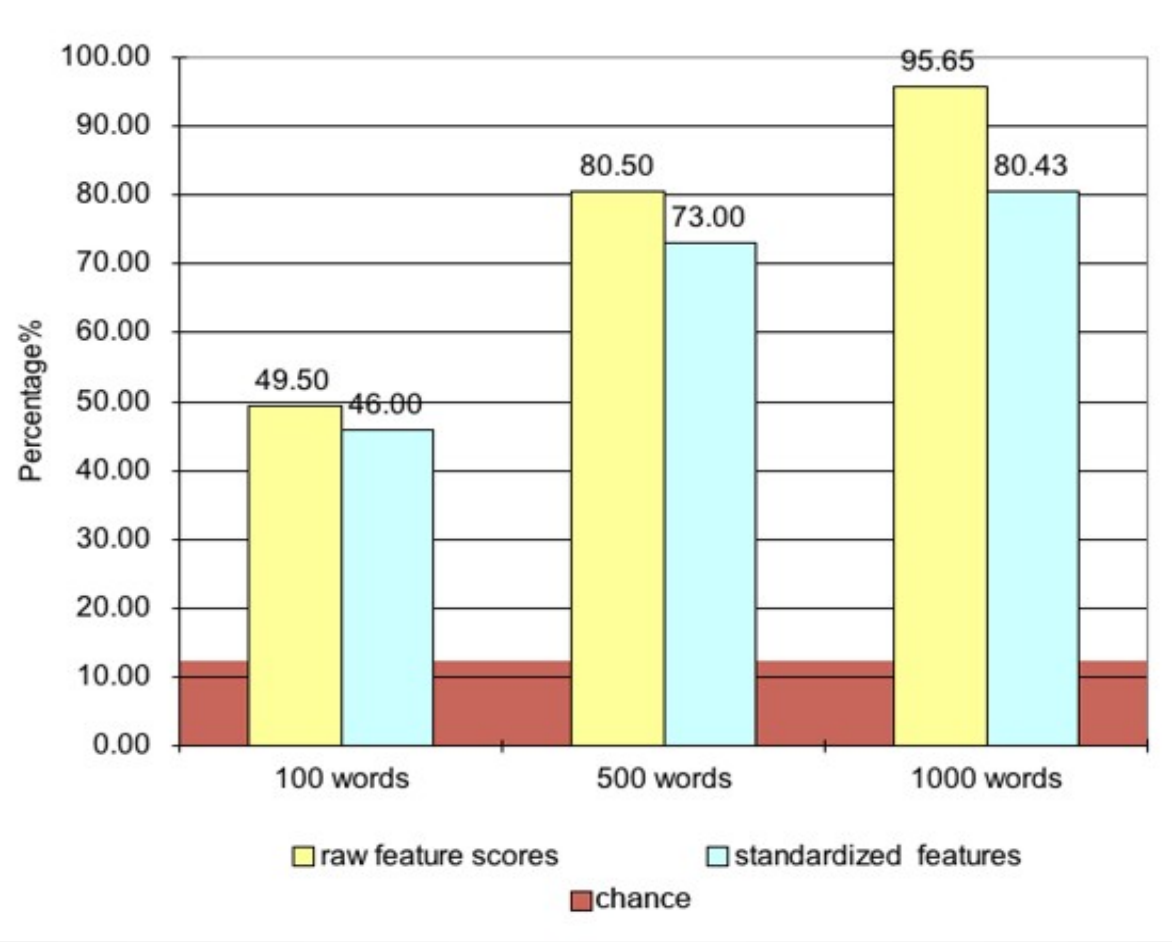
- Authorship anomalies
  - 8 Victorian authors
- Factual writing vs. opinion writing anomalies
  - Opinion columns
- Subversive article anomalies
  - Newswire vs. Anarchist Cookbook
- Machine translation anomalies
  - Chinese news translated into English

Authorship Anomalies



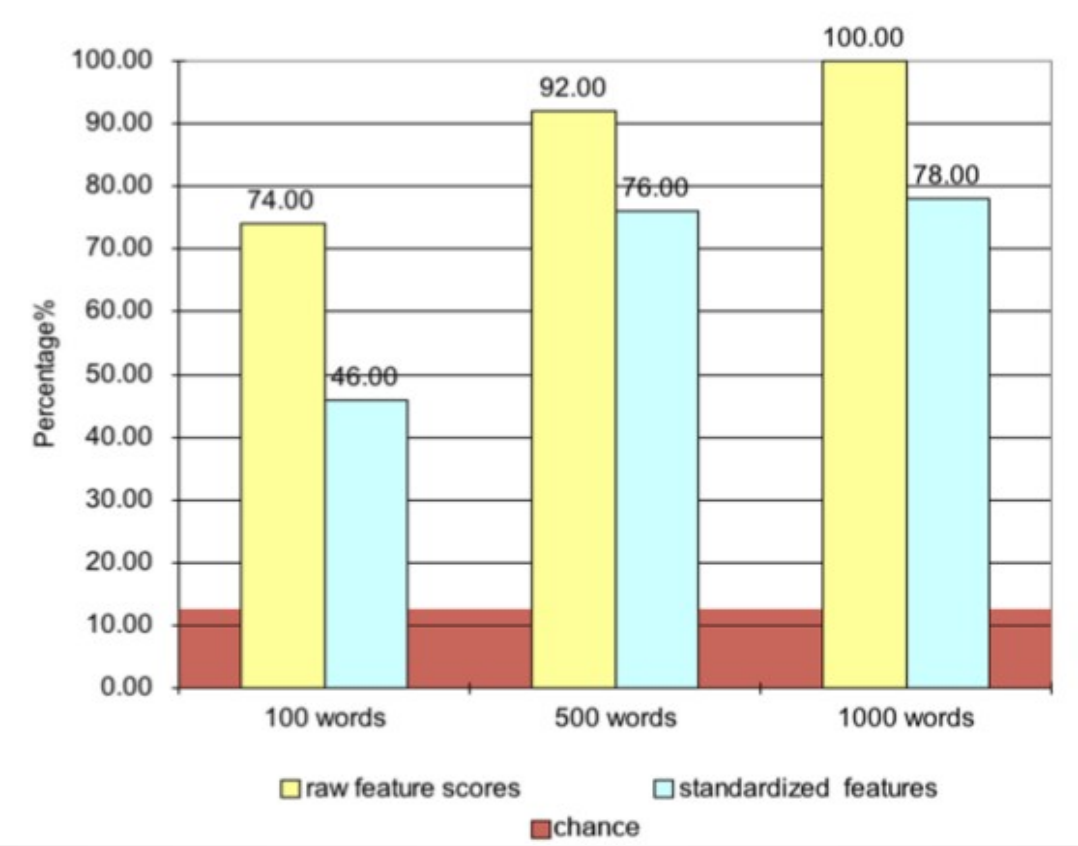
Fact vs. Opinion

- Opinion (editorials, opinion columns) added into factual article (Gigaword Corpus)



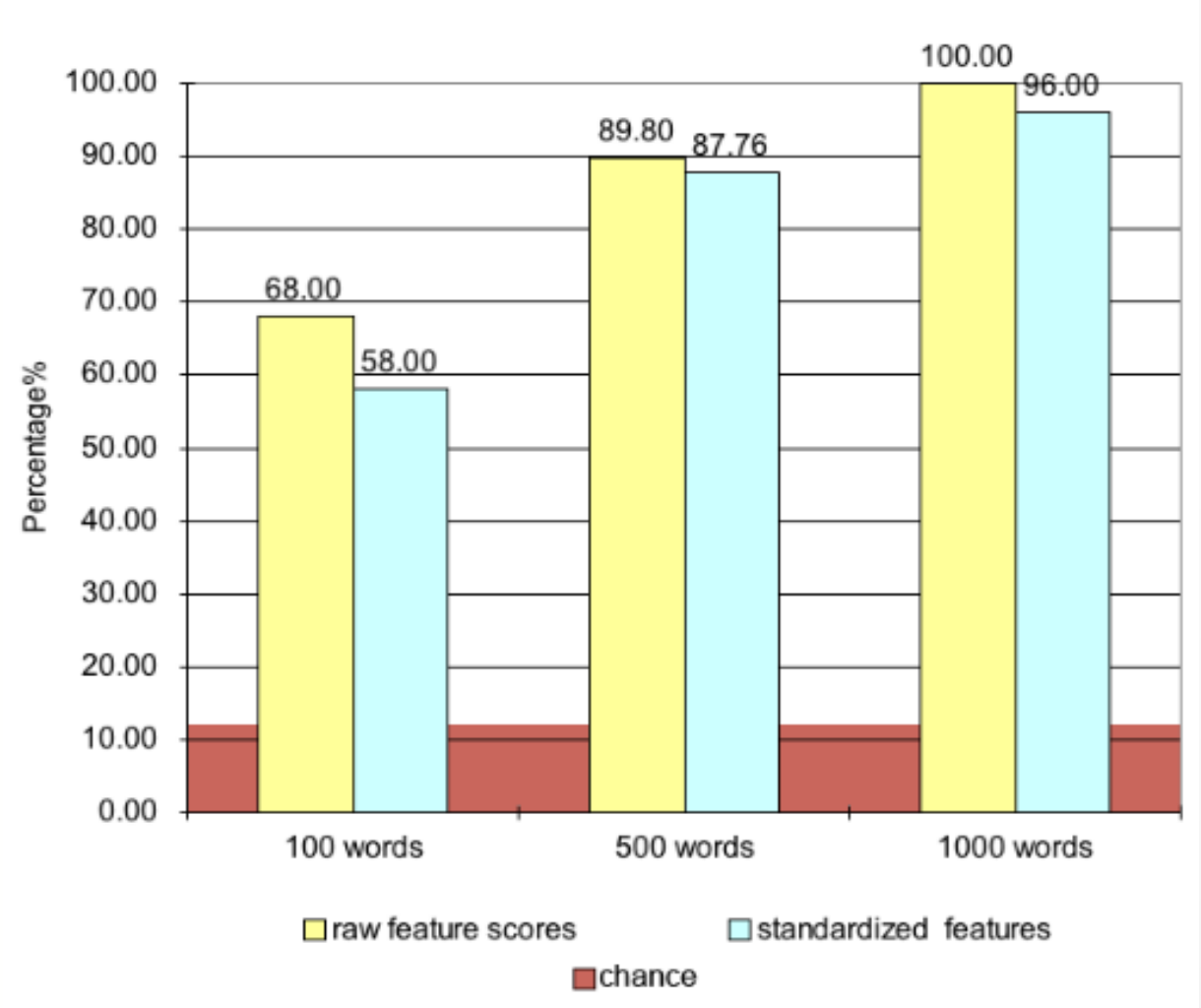
Newswire vs. Anarchist Cookbook

- Anarchist Cookbook – recipes of explosives, instructions how to build different devices, …
- Difference in genre



Machine Translation Anomalies

- Usage of Google Translate

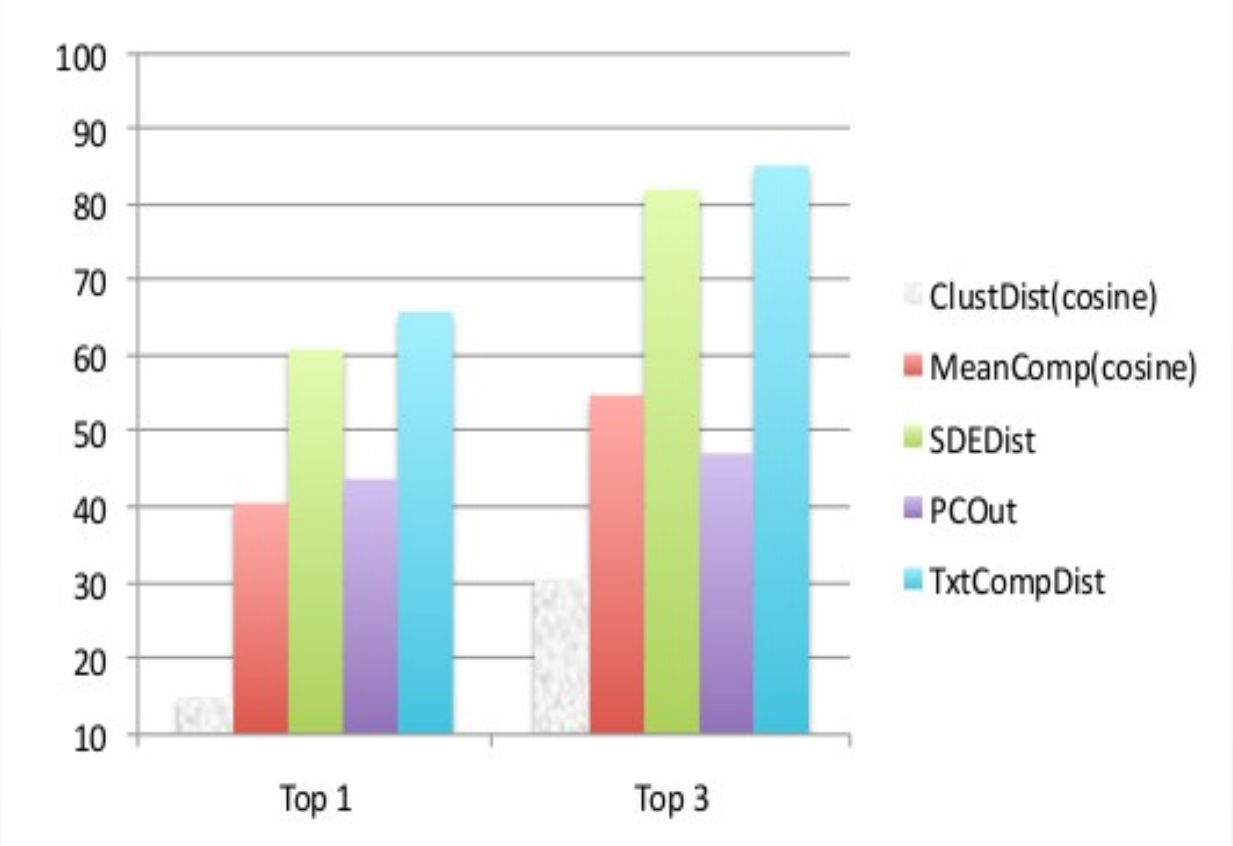


Conclusions from Experiments [1]

- Best results for detecting anomalies based on difference in genre or style
  - Identification of machine translated text
  - Newswire vs. Anarchist Cookbook
- Difficult to identify anomaly in Top 1
  - 96% probability for MT task (large segment)
  - 2% probability by chance

Conclusions from Experiments [2]

- Best results for TxtCompDist
- SDEDist – higher Time costs



Precision & Recall

- What is probability that given segment is anomalous?
- Definition of threshold for anomality score
  - Maximal precision (100%)
  - Best recall possible

Precision & Recall [2]

Recall/Precision (Threshold)	Segment Size	Chinese Translations	Fact vs Opinion	Anarchists Cookbook	all
	100	52%/100% (369)	46%/100% (369)	36%/100% (371)	44%/100% (371)
	500	83%/100% (279)	38%/100% (280)	66%/100% (280)	62%/100% (280)
	1000	96%/100% (252)	43%/100% (252)	88%/100% (252)	76%/100% (252)
	all	46.3%/100% (370)	22.2%/100% (370)	30%/100% (370)	33%/100% (370)

Feature Selection

- Which features help us to identify an anomaly and which don't
- Score: difference in values for anomalous segments and normal segments
- Results
  - Least effective are emotional features
  - Most effective features are basically the same for all anomality types, least effective features differ

Most Effective Features

- Gunning-Fog Index
- Percentage of passive sentences
- Flesch-Kincaid Reading Ease
- Percentage of sentences longer than 15 words
- Lix Formula

Least Effective Features

- Words of economic, commetcial, industrial orientation
- Terms denoting Kinship
- Words for non-work social rituals
- Words concerned with fetching or carrying
- Words for places occuring in nature

Summary of Conclusions

- Variations in text viewed as outliers
- Best method: comparing segments with its textual complement
- With larger segment increases accuracy
- The easiest anomalies to identify are anomalies in genre or style
- Usefulness of stylistic features, word distributions, different readibility measures, …

IS and MUNI Applications

- Thesis and essays plagiarism
- Discussion Forum
- Log Entries

Thesis and Essays Plagiarism

- Looking for segments from external sources
- Authorship anomalies
- Identification of machine translated text
  - Does not have to be plagiarism

Discussion Forums

- Looking for irrelevant posts
- Problems with data
  - Sentence boundaries
  - Special symbols (math, chemistry,...)
  - Length varies
- Anomalies in content, not in style

Log Entries

- e.g. behaviour of students
- Stream processing
- Many similar entries
- Short length of entries

Thank You for Your Attention
Questions?

Appendix [1] - Authorship

Top n Segments	Percentage of the time found	Percentage of the time found (standardized features)	Chance
Segment size: 100 words			
1	13.54%	16.25%	1.96%
3	25.54%	31.25%	6.00%
5	32.61%	40.46%	10.21%
10	48.57%	52.04%	21.59%
20	63.79%	67.82%	49.16%
Segment size: 500 words			
1	29.01%	37.79%	1.96%
3	46.04%	50.72%	6.00%
5	49.53%	60.59%	10.21%
10	61.67%	72.40%	21.59%
20	74.30%	83.88%	49.16%
Segment size: 1000 words			
1	44.80%	48.02%	1.96%
3	54.98%	66.60%	6.00%
5	60.08%	74.07%	10.21%
10	76.26%	85.79%	21.59%
20	96.19%	97.88%	49.16%

Appendix [2] – Fact vs. Opinion

Top n Segments	Percentage of the time found	Percentage of the time found (standardized features)	Chance
Segment size: 100 words			
1	26.50%	17.50%	1.96%
3	46.00%	36.00%	6.00%
5	49.50%	46.00%	10.21%
10	62.00%	64.00%	21.59%
20	78.50%	76.00%	49.16%
Segment size: 500 words			
1	13.50%	22.00%	1.96%
3	50.50%	59.00%	6.00%
5	80.50%	73.00%	10.21%
10	90.50%	82.50%	21.59%
20	99.00%	96.00%	49.16%
Segment size: 1000 words			
1	34.78%	53.26%	1.96%
3	85.87%	73.91%	6.00%
5	95.65%	80.43%	10.21%
10	98.91%	94.57%	21.59%
20	98.91%	98.91%	49.16%

Appendix [3] – Anarchist Cookbook

Top n Segments	Percentage of the time found	Percentage of the time found (standardized features)	Chance
Segment size: 100 words			
1	38.00%	34.00%	1.96%
3	68.00%	38.00%	6.00%
5	74.00%	46.00%	10.21%
10	88.00%	58.00%	21.59%
20	98.00%	82.00%	49.16%
Segment size: 500 words			
1	70.00%	24.00%	1.96%
3	90.00%	58.00%	6.00%
5	92.00%	76.00%	10.21%
10	100.00%	78.00%	21.59%
20	100.00%	100.00%	49.16%
Segment size: 1000 words			
1	88.78%	36.26%	1.96%
3	100.00%	58.00%	6.00%
5	100.00%	78.00%	10.21%
10	100.00%	94.00%	21.59%
20	100.00%	98.00%	49.16%

Appendix [4] – MT

Top n Segments	Percentage of the time found	Percentage of the time found (standardized features)	Chance
Segment size: 100 words			
1	54.00%	36.00%	1.96%
3	60.00%	54.00%	6.00%
5	68.00%	58.00%	10.21%
10	74.00%	60.00%	21.59%
20	80.00%	76.00%	49.16%
Segment size: 500 words			
1	83.67%	59.18%	1.96%
3	87.76%	69.39%	6.00%
5	89.80%	87.76%	10.21%
10	93.88%	93.88%	21.59%
20	100.00%	100.00%	49.16%
Segment size: 1000 words			
1	96.00%	92.00%	1.96%
3	100.00%	96.00%	6.00%
5	100.00%	96.00%	10.21%
10	100.00%	100.00%	21.59%
20	100.00%	100.00%	49.16%