Same but Different: Distant Supervision for Predicting and Understanding Entity Linking Difficulty

Renato Stoffalette João[1], Pavlos Fafalios[1], Stefan Dietze[1,2]
joao@L3S.de

[1]L3S Research Center / Leibniz University of Hanover
Hannover, Germany
[2]GESIS – Leibniz Institute for the Social Sciences
Köln, Germany

April, 2019

Introduction

- Entity Linking (EL) or Named Entity Recognition and Disambiguation (NERD) is the task of recognizing entity mentions in text and link them to an entity in a reference knowledge base
    - Web search, IR, docs classification, etc

Introduction

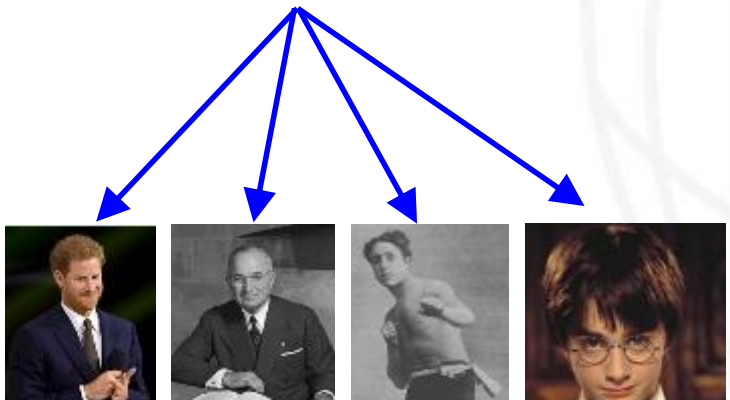
- Entity Linking (EL) or Named Entity Recognition and Disambiguation (NERD) is the task of recognizing entity mentions in text and link them to an entity in a reference knowledge base
    - Web search, IR, docs classification, etc

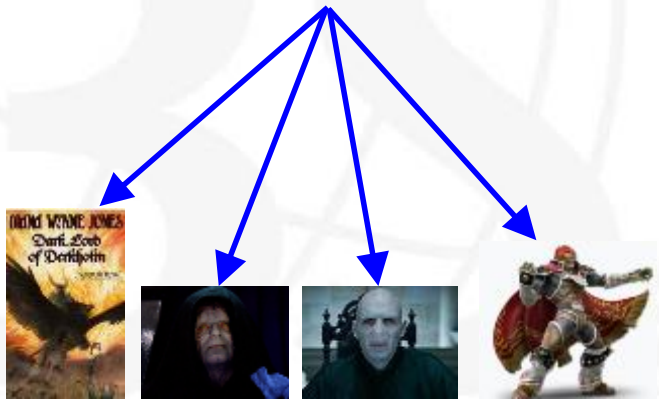
Harry fought with you know who. He defeated the dark lord.

Introduction

- Entity Linking (EL) or Named Entity Recognition and Disambiguation (NERD) is the task of recognizing entity mentions in text and link them to an entity in a reference knowledge base
  - Web search, IR, docs classification, etc

[Harry] fought with you know who. He defeated [the dark lord.]

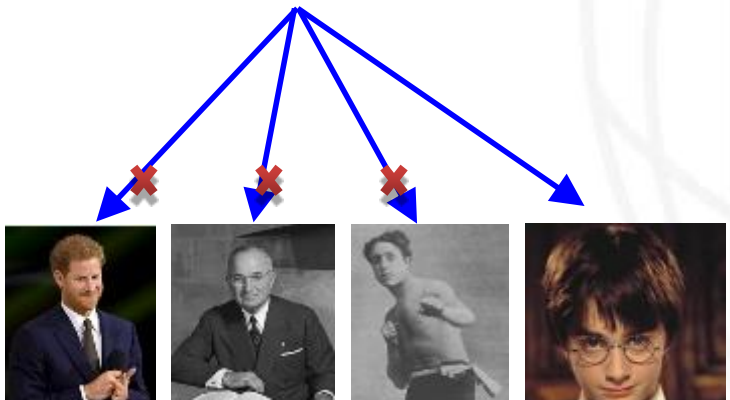


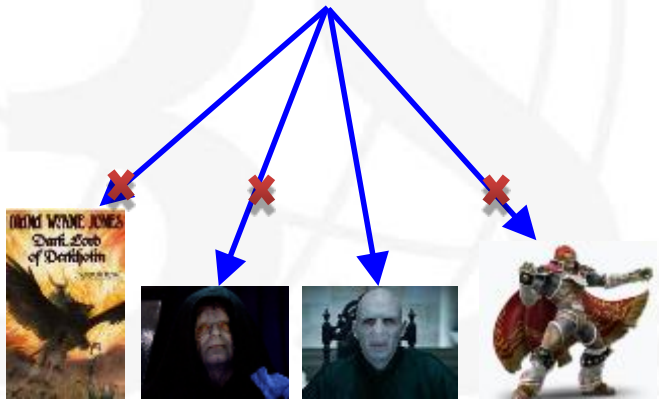


Introduction

- Entity Linking (EL) or Named Entity Recognition and Disambiguation (NERD) is the task of recognizing entity mentions in text and link them to an entity in a reference knowledge base
  - Web search, IR, docs classification, etc

[Harry] fought with you know who. He defeated [the dark lord.]



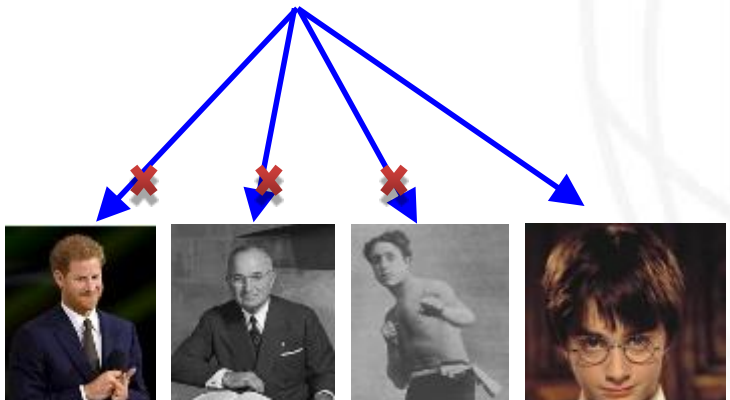


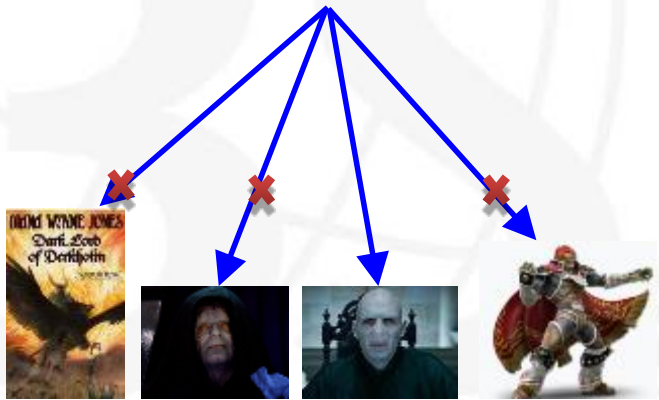
Introduction

- Entity Linking (EL) or Named Entity Recognition and Disambiguation (NERD) is the task of recognizing entity mentions in text and link them to an entity in a reference knowledge base
  - Web search, IR, docs classification, etc

High P and R are required if EL is to have a positive impact in applications!

[Harry] fought with you know who. He defeated [the dark lord.]





Motivation

- Gerbil benchmark [1] has shown EL systems' performance is highly affected by the characteristics of the data sets
  - Number of entities per document
  - Document length
  - Total number of entities
  - Salient entity types

- Applying state-of-the-art EL systems out of the box does not provide the best performance

[1] Röder, M., Usbeck, R., & Ngonga Ngomo, A. C. (2017). Gerbil–benchmarking named entity recognition and linking consistently. Semantic Web, (Preprint), 1-21.

Motivation

- EL difficulty varies per corpus but also with each individual mention
- Difficult to link mentions often share common characteristics
  - Highly ambiguous mentions with large number of candidates
    - *"Brown"*, *"Smith"*, *"Williams"*
  - Mentions of long-tail entities
    - *A local deputy, local team player, reporter, etc*
  - Mentions of entities where the respective meaning evolves significantly over time
    - *"President of the US"*, *"the Pope"*
  - Mentions of entities where the popularity changes significantly over time
    - *"Amazon"* in 1980 or 2018, *"Watson"* in 1990 or 2015

Motivation

- EL systems never reach perfect P/R on arbitrary corpus
  - Human judgments can be incorporated into the pipeline to improve EL results
- Estimating a priori the difficulty of linking a particular mention can facilitate high P/R systems
  - e.g. Flagging critical mentions which require manual judgments

Related Works

- *Shen et. al.*[2] presents overview of main EL approaches
- Diverse EL systems
  - News documents, tweets, queries, web lists, etc
  - Medicine, music domain, scientific publications,etc
- Length and num. of candidate entities affect EL difficulty Hoffart *et. al.*[3]
  - KORE50: large number of candidate entities
  - WP: short (and thus very ambiguous) mentions

[2] Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2), 443-460.

[3] Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., & Weikum, G. (2012, October). KORE: keyphrase overlap relatedness for entity disambiguation. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 545-554). ACM.

Contributions

- An automated approach to generate difficulty labels
  - Based on agreement/disagreement among EL systems
  - Labels can be used to improve semi-automated EL pipelines

- Novel approach, features sets and classifiers for predicting EL difficulty
  - Detect latent, corpus specific characteristics that affect EL performance

- Evaluation results
  - Effectiveness on predicting and understanding EL difficulty
  - Effectiveness on improving semi-automated EL pipelines

Consensus-based Labelling

Consensus Based Labelling

Problem Formulation

Predict the difficulty in linking a mention **m** to an entity **e** in a knowledge base as a multiclass classification problem where **m** is assigned to one of the following classes

- ***HARD*** **-** EL systems usually fail to find the correct link
- ***EASY*** **-** EL systems almost always find the correct link
- ***MEDIUM*** - All other cases (neither EASY nor HARD)

Consensus Based Labelling

Labelling Process

- Instead of costly manual labelling
- Automated approximation strategy with SOTA EL systems
- EL systems agreement is used as indicator
  - ***HARD*** **-** All EL systems disagree on provided link
  - ***EASY*** **-** All EL system agree on provided link
  - ***MEDIUM*** - All other cases (neither **EASY** nor **HARD**)

Consensus Based Labelling

Labelling Process Limitations

- Assumption the provided link is correct
  - False Positives (EASY cases)
- Requires mentions to be recognized by all the EL systems

Supervised Classification can be used to predict EL difficulty

Learning Entity Linking Difficulty

Learning Entity Linking difficulty

- Distantly supervised classification
  - Trained using the proposed labelling strategy
  - Predict the linking difficulty of arbitrary entity mentions

- A diverse set of features is needed
  - *What characterizes a difficult to link mention?*

Learning Entity Linking difficulty

Category	Notation	Description
Mention	m_{len}	Num of mention's characters (length).
	m_{words}	Num of mention's words.
	m_{freq}	Num of mention's occurrences in the doc (frequency).
	m_{df}	Num of docs in the corpus containing at least one occurrence of the mention.
	m_{cand}	Num of mention's candidate entities in a reference KB.
	m_{pos}	Mention's normalised position in the doc (num of chars from the start of the doc / total num of doc's chars).
	m_{sent}	Num of chars of the sentence containing the mention.
Document	d_{words}	Num of words in the document containing the mention.
	d_{topic}	Main topic discussed in the document containing the mention (e.g., SPORTS, or POLITICS).
	d_{ents}	Num of entity mentions recognised in the document containing the mention.
Temporal	t_{age}	The distance (age) of the doc's publication date from the date of the reference KB.
	t_{df}	Number of docs containing at least one occurrence of the mention, published within k intervals from the doc's publication date (e.g., $+/-$ 6 months).
	t_{jmin} / t_{jmax} / t_{javg}	Min, max and avg Jaccard similarity of the mention's top-K similar words (computed using Word2Vec) for all pairs of consecutive time periods of fixed granularity.

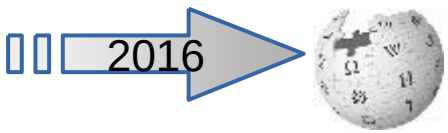
Experimental Evaluation

Experimental Evaluation

- New York Times (NYT) Annotated Corpus
- 1.8 million articles published between Jan. 1987 – Jun. 2007
- Range of topics (sports, politics, local news, arts, business, technology, etc.)
- Diverse formats (long texts, short notices, corrections, and headlines)
- Number of articles per year ranges from 79,077 (in 2007) to 106,104 (in 1987)

Labelling

- We applied the proposed labelling strategy using three widely-used EL systems
  - Ambiverse (previously AIDA) [Hoffart et al., 2011]
  - Babelfy [Moro et al., 2014]
  - TagMe [Ferragina and Scaiella, 2010]



System	#Annotations	#Dist. mentions	#Dist. entities
Ambiverse	45,807,845	817,957	721,065
Babelfy	109,360,676	1,044,940	1,117,829
TagMe	55,907,818	1,474,772	1,002,874

Quality of the generated Labels

- **HARD** - manual evaluation using a random sample of 500 mentions
  - Ambiverse - 24%
  - Babelfy - 16%
  - Tagme - 31 %
- **EASY** - manual evaluation using a random sample of 200 mentions
  - 95% accuracy
- **MEDIUM** - manual evaluation using a random sample of 200 mentions
  - Test if the two systems that agree provide the correct entity
  - 88% accuracy

Quality of the generated Labels

- **Original Class imbalance**
  - **HARD -** 2.9%
  - **EASY -** 78.6%
  - **MEDIUM** – 21.4 %
- *Considering the original imbalance distribution (majority of cases are **EASY**)*
  - Expected error rate of **MEDIUM** and **EASY** label: < 7%

Sampling and Balancing

- **Original Class imbalance**
  - **HARD** - 2.9%
  - **EASY** - 78.6%
  - **MEDIUM** – 21.4 %

- **UNBALANCED** : Maintaining the actual class distribution as observed in the data

- **BALANCED** : Random undersampling of the majority class/classes (all classes have the same number of training instances)
  - 10-fold cross validation
  - Test set maintains original class distribution

Sampling and Balancing

- **SAMPLE25** - Random 25% stratified sample of the full dataset
- **SAMPLE10** - Random 10% stratified sample of the full dataset
- **SAMPLE1** - Random 1% stratified sample of the full dataset

Dataset	#Instances	#HARD	#MEDIUM	#EASY
SAMPLE25	2,969,108	85,059	616,420	2,267,629
SAMPLE10	1,187,642	34,023	246,568	907,051
SAMPLE1	118,763	3,402	24,656	90,705

Classification Models

- Naive Bayes
- Logistic Regression
- Decision Tree
- Random Forest

Baselines

Ambiguity is strongly dependent on the candidates available in a KB as well as the mention's length [4]

- **CANDIDNUM** : Classification using only the feature m_{cand} (num of mention's candidate entities in the reference KB)

- **MENTLENGTH** : Classification using only the feature m_{len} (mention's length)

[4] Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., & Weikum, G. (2012, October). KORE: keyphrase overlap relatedness for entity disambiguation. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 545-554). ACM.

Evaluation Measures

- ***Precision*** - the fraction of the correctly classified instances among the instances assigned to the class
- ***Recall*** - the fraction of the correctly classified instances among all instances of the class
- F_1 - The harmonic mean of recall and precision

- Per class and the macro average performance
  - To ensure the size of each class has no impact on the representativeness of our metrics

Classification Performance

- Overall prediction performance (macro average) using SAMPLE25

Method	Model	Unbalanced			Balanced		
		P	R	F1	P	R	F1
CandidNum	Naive Bayes	0.38	0.35	0.34	0.37	0.40	0.32
	Logistic Regr.	0.31	0.33	0.30	0.43	0.41	0.35
	Decision Tree	0.74	0.47	0.50	0.48	0.61	0.47
	Random Forest	0.74	0.47	0.50	0.48	0.61	0.47
MentLength	Naive Bayes	0.25	0.33	0.29	0.36	0.42	0.26
	Logistic Regr.	0.25	0.33	0.29	0.37	0.44	0.31
	Decision Tree	0.25	0.33	0.29	0.42	0.47	0.40
	Random Forest	0.25	0.33	0.29	0.42	0.47	0.39
MultiFeature	Naive Bayes	0.42	0.41	0.41	0.43	0.49	0.41
	Logistic Regr.	0.45	0.36	0.35	0.43	0.50	0.40
	Decision Tree	0.74	0.69	0.71	0.56	0.74	0.59
	Random Forest	**0.83**	**0.72**	**0.76**	**0.58**	**0.76**	**0.60**

Classification Performance

- Overall prediction performance (macro average) using SAMPLE25

Method	Model	Unbalanced			Balanced		
		P	R	F1	P	R	F1
CandidNum	Naive Bayes	0.38	0.35	0.34	0.37	0.40	0.32
	Logistic Regr.	0.31	0.33	0.30	0.43	0.41	0.35
	Decision Tree	0.74	0.47	0.50	0.48	0.61	0.47
	Random Forest	0.74	0.47	0.50	0.48	0.61	0.47
MentLength	Naive Bayes	0.25	0.33	0.29	0.36	0.42	0.26
	Logistic Regr.	0.25	0.33	0.29	0.37	0.44	0.31
	Decision Tree	0.25	0.33	0.29	0.42	0.47	0.40
	Random Forest	0.25	0.33	0.29	0.42	0.47	0.39
MultiFeature	Naive Bayes	0.42	0.41	0.41	0.43	0.49	0.41
	Logistic Regr.	0.45	0.36	0.35	0.43	0.50	0.40
	Decision Tree	0.74	0.69	0.71	0.56	0.74	0.59
	Random Forest	**0.83**	**0.72**	**0.76**	**0.58**	**0.76**	**0.60**

Classification Performance

- Overall prediction performance (macro average) using SAMPLE25

Method	Model	Unbalanced			Balanced		
		P	R	F1	P	R	F1
CandidNum	Naive Bayes	0.38	0.35	0.34	0.37	0.40	0.32
	Logistic Regr.	0.31	0.33	0.30	0.43	0.41	0.35
	Decision Tree	0.74	0.47	0.50	0.48	0.61	0.47
	Random Forest	0.74	0.47	0.50	0.48	0.61	0.47
MentLength	Naive Bayes	0.25	0.33	0.29	0.36	0.42	0.26
	Logistic Regr.	0.25	0.33	0.29	0.37	0.44	0.31
	Decision Tree	0.25	0.33	0.29	0.42	0.47	0.40
	Random Forest	0.25	0.33	0.29	0.42	0.47	0.39
MultiFeature	Naive Bayes	0.42	0.41	0.41	0.43	0.49	0.41
	Logistic Regr.	0.45	0.36	0.35	0.43	0.50	0.40
	Decision Tree	0.74	0.69	0.71	0.56	0.74	0.59
	Random Forest	**0.83**	**0.72**	**0.76**	**0.58**	**0.76**	**0.60**

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	**0.75**	**0.46**	**0.57**	**0.83**	**0.71**	**0.77**	**0.92**	0.97	**0.95**	**0.21**	**0.84**	**0.34**	**0.57**	**0.63**	**0.60**	**0.95**	**0.82**	**0.88**

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	**0.75**	**0.46**	**0.57**	**0.83**	**0.71**	**0.77**	**0.92**	0.97	**0.95**	**0.21**	**0.84**	**0.34**	**0.57**	**0.63**	**0.60**	**0.95**	**0.82**	**0.88**

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	**0.75**	**0.46**	**0.57**	**0.83**	**0.71**	**0.77**	0.92	0.97	**0.95**	**0.21**	**0.84**	**0.34**	**0.57**	**0.63**	**0.60**	0.95	**0.82**	**0.88**

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	0.75	0.46	0.57	0.83	0.71	0.77	0.92	0.97	0.95	0.21	0.84	0.34	0.57	0.63	0.60	0.95	0.82	0.88

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	0.75	0.46	0.57	0.83	0.71	0.77	0.92	0.97	0.95	0.21	0.84	0.34	0.57	0.63	0.60	0.95	0.82	0.88

Classification Performance

- Per class prediction performance using SAMPLE25

Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	0.75	0.46	0.57	**0.83**	**0.71**	**0.77**	**0.92**	0.97	**0.95**	0.21	0.84	0.34	**0.57**	**0.63**	**0.60**	**0.95**	**0.82**	**0.88**

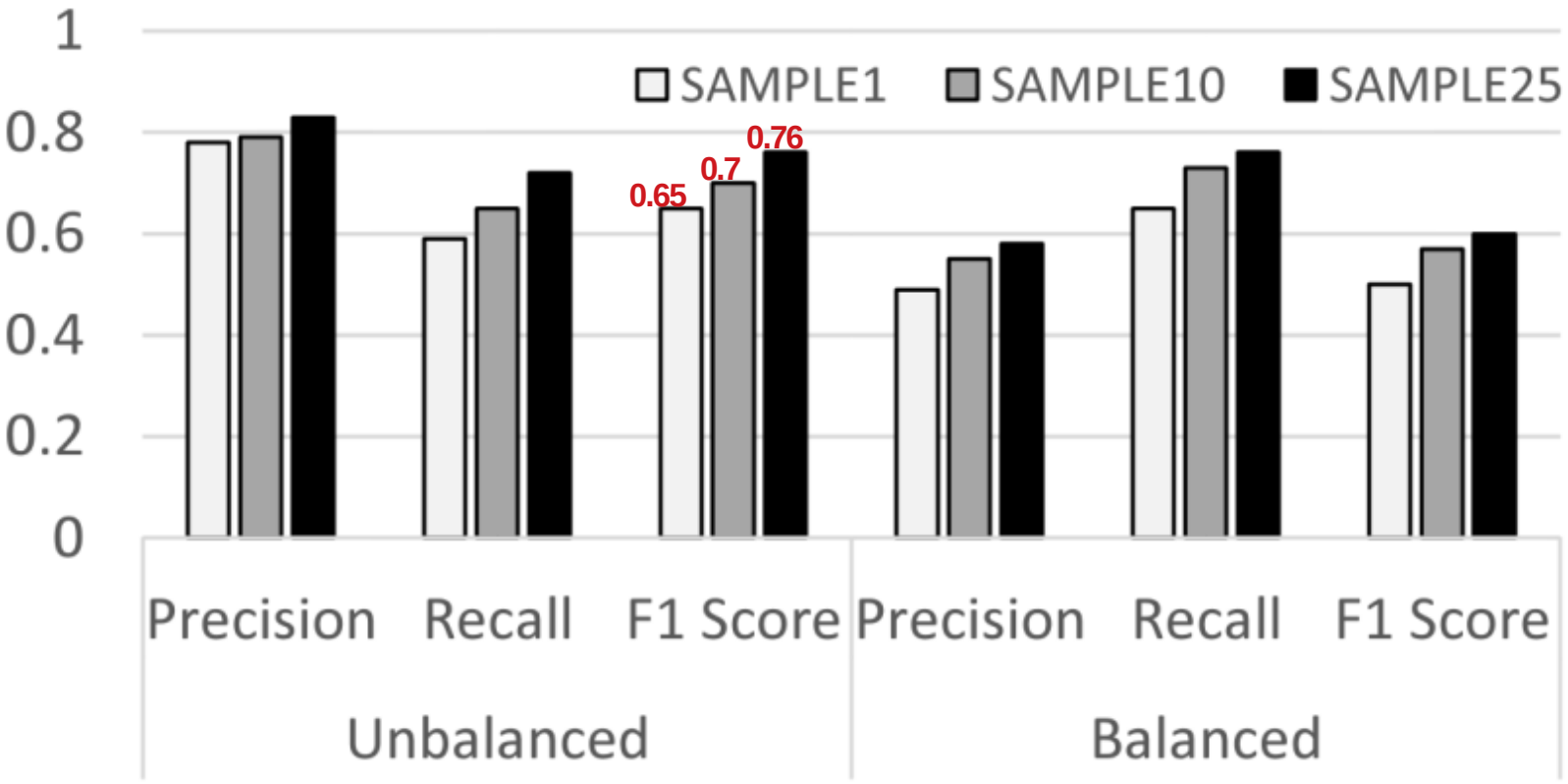
Classification Performance

- Per class prediction performance using SAMPLE25

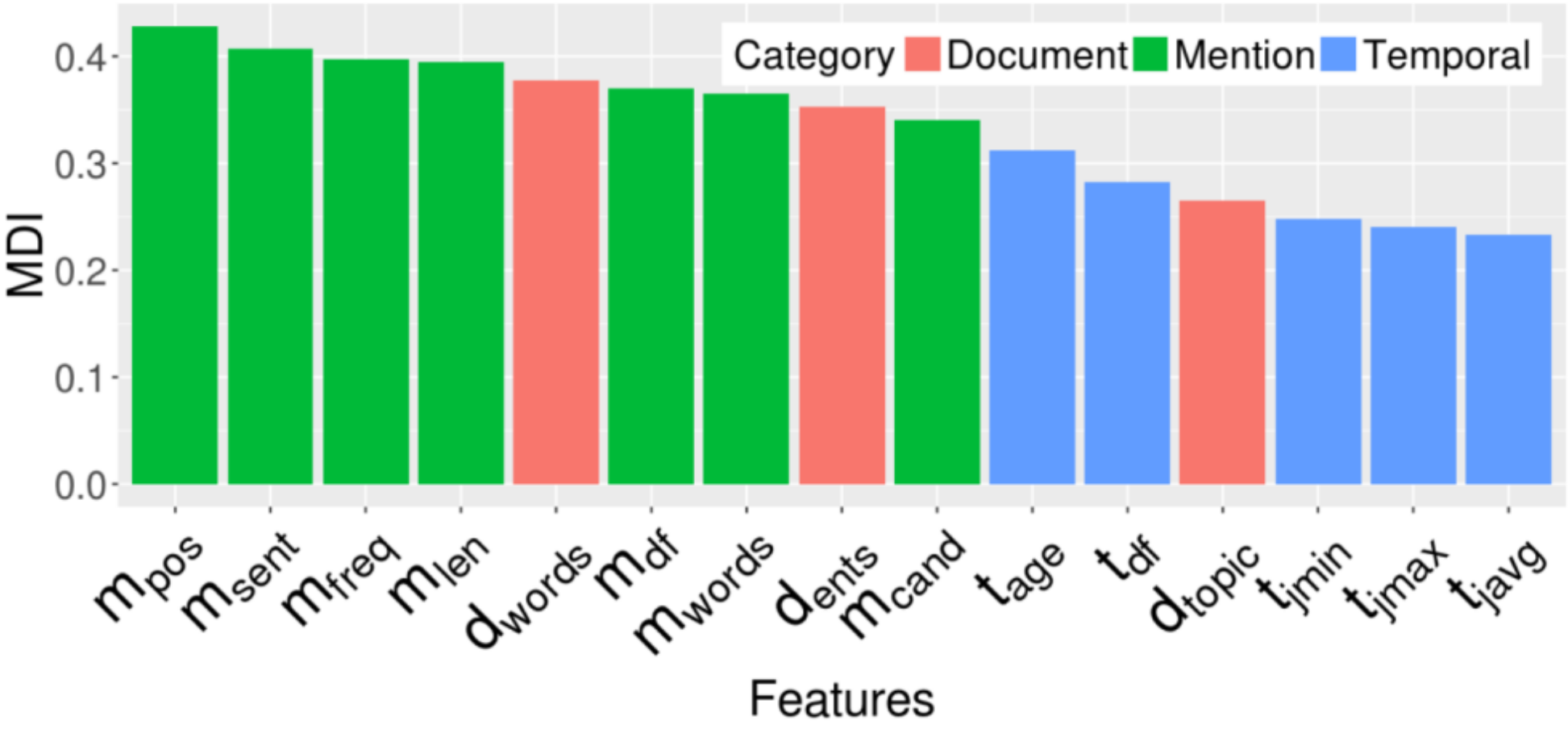
Method	Model	Unbalanced Training									Balanced Training								
		Hard			Medium			Easy			Hard			Medium			Easy		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CandidNum	Naive Bayes	0	0	0	0.37	0.11	0.17	0.78	0.96	0.86	0.06	0.32	0.10	0.26	0.03	0.05	0.80	0.86	0.83
	Logistic Regr.	-	0	-	0.16	0.02	0.04	0.76	0.97	0.85	0.05	0.33	0.09	0.41	0.09	0.14	0.82	0.83	0.82
	Decision Tree	0.67	0.08	0.14	0.73	0.35	0.47	0.83	0.98	0.90	0.10	0.65	0.17	0.43	0.48	0.45	0.92	0.69	0.79
	Random Forest	0.67	0.08	0.14	0.72	0.35	0.47	0.83	0.97	0.90	0.09	0.66	0.16	0.44	0.47	0.45	0.92	0.69	0.79
MentLength	Naive Bayes	-	0	-	-	0	-	0.76	1	0.87	0.05	0.72	0.08	0.14	0.12	0.13	0.88	0.43	0.58
	Logistic Regr.	-	0	-	-	0	-	0.76	1	0.87	0.05	0.61	0.09	0.18	0.13	0.15	0.87	0.57	0.69
	Decision Tree	-	0	-	-	0	-	0.76	1	0.87	0.06	0.40	0.10	0.33	0.38	0.35	0.87	0.64	0.74
	Random Forest	-	0	-	-	0	-	0.76	1	0.87	0.06	0.41	0.10	0.33	0.38	0.35	0.87	0.63	0.73
MultiFeature	Naive Bayes	0.04	0.01	0.02	0.40	0.39	0.39	0.82	0.84	0.83	0.06	0.47	0.11	0.37	0.31	0.34	0.86	0.70	0.77
	Logistic Regr.	0	0	0	0.58	0.11	0.18	0.78	**0.98**	0.87	0.07	0.49	0.12	0.33	0.39	0.36	0.88	0.63	0.74
	Decision Tree	0.55	0.43	0.48	0.76	0.69	0.73	**0.92**	0.95	0.94	0.20	0.79	0.32	0.55	0.62	0.58	**0.95**	0.81	0.87
	Random Forest	0.75	0.46	0.57	**0.83**	**0.71**	**0.77**	**0.92**	0.97	**0.95**	0.21	0.84	0.34	**0.57**	**0.63**	**0.60**	**0.95**	**0.82**	**0.88**

Classification Performance

- Influence of dataset size on prediction performance (**macro average**) (**Random Forest** classifier)

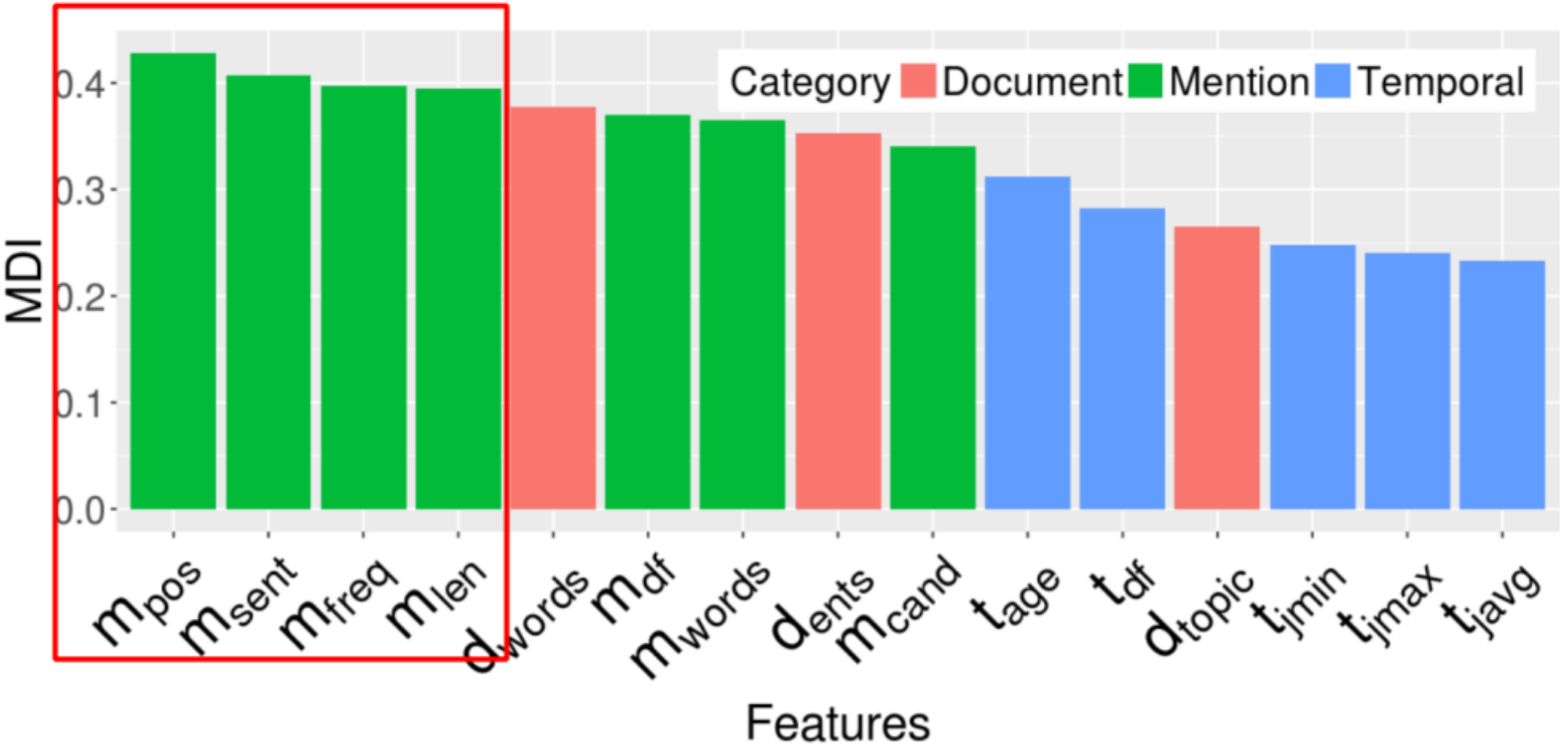


Feature Analysis



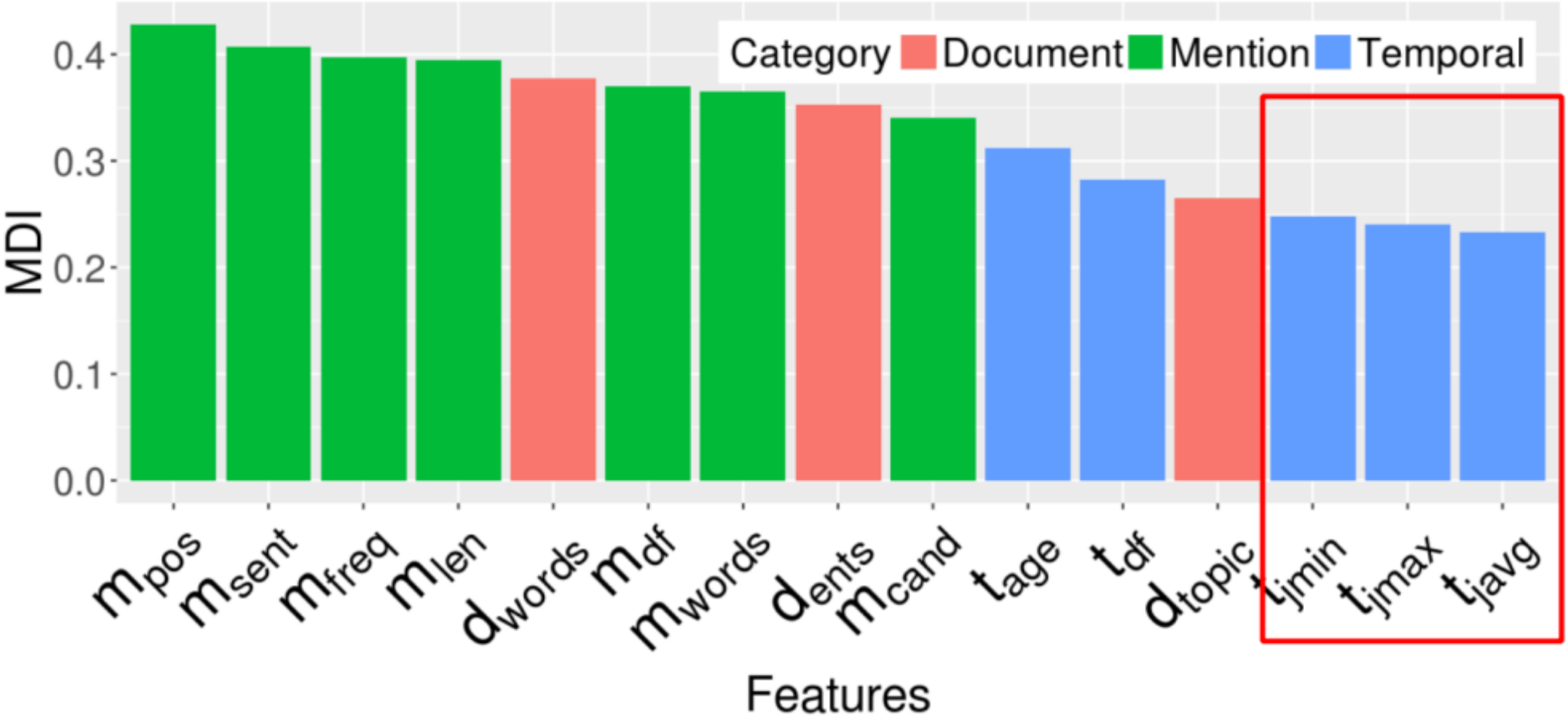
- Mean Decrease in Impurity (**MDI**) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits.

Feature Analysis



- Mean Decrease in Impurity (**MDI**) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits.

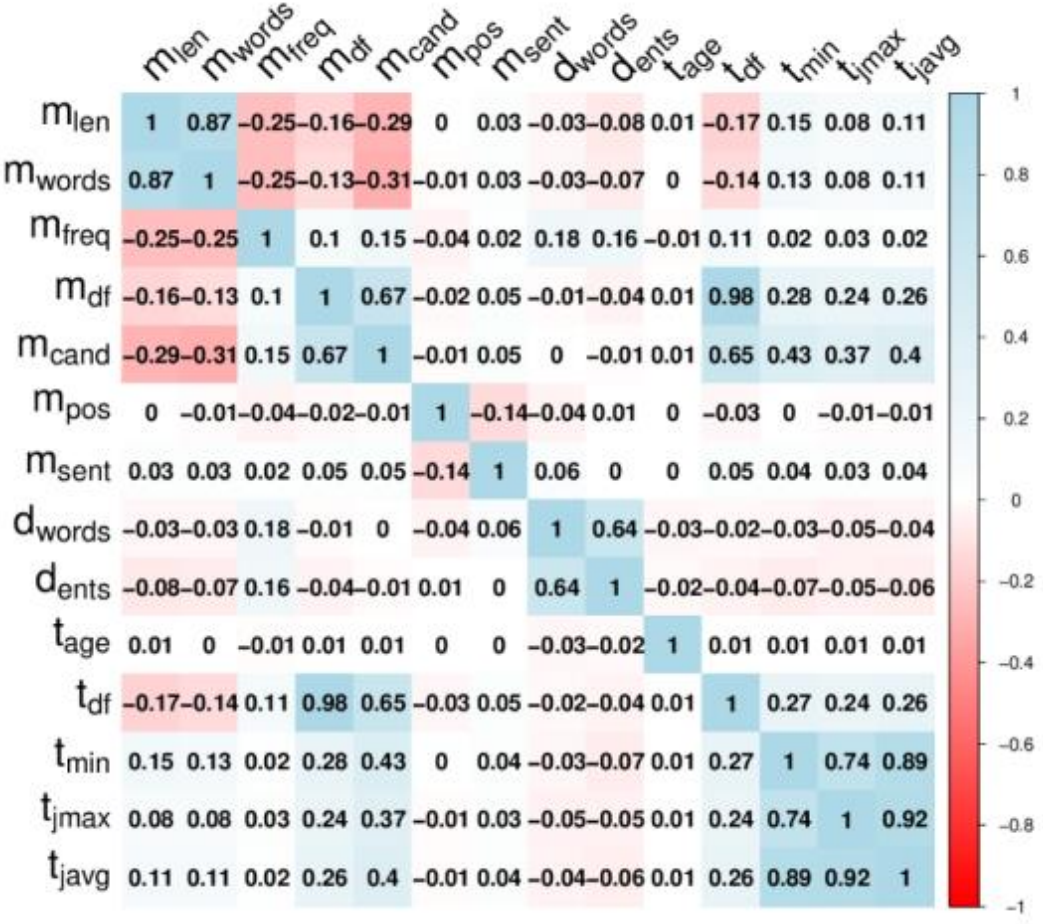
Feature Analysis



- Mean Decrease in Impurity (**MDI**) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits.

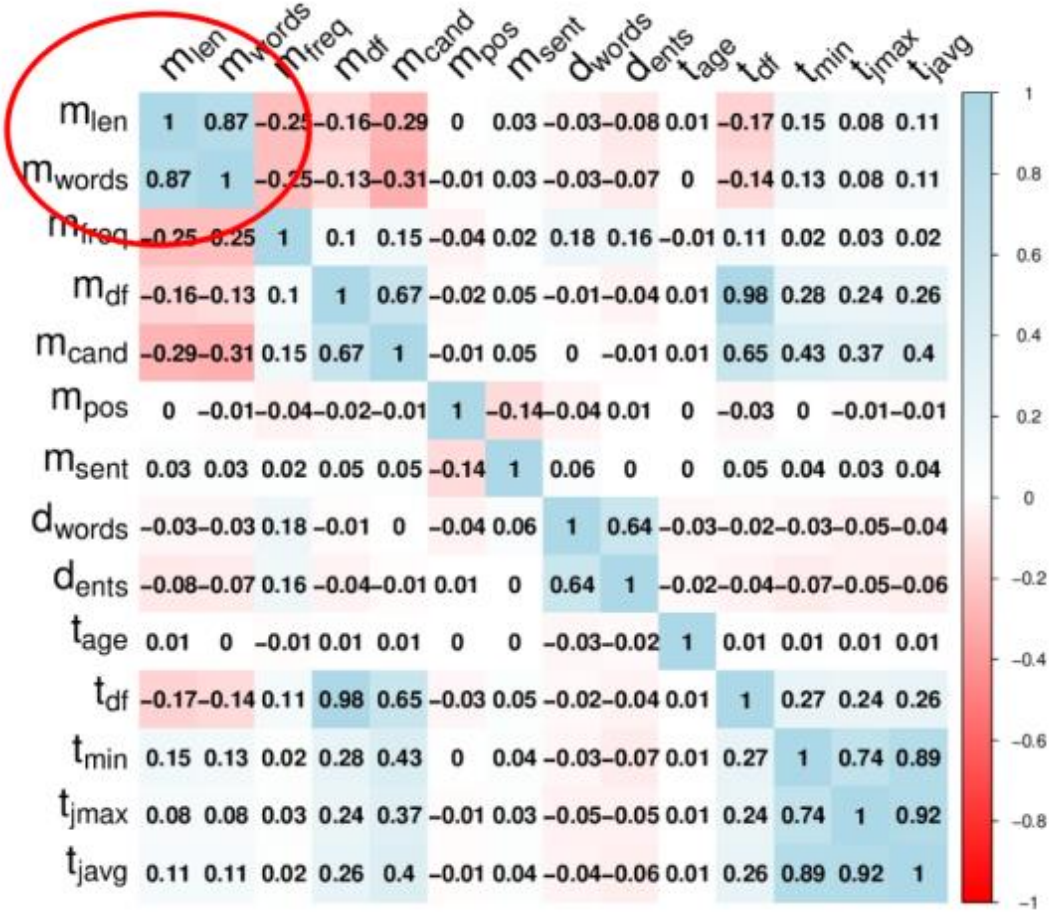
Feature Analysis

Correlation among features – **Pearson's ρ**



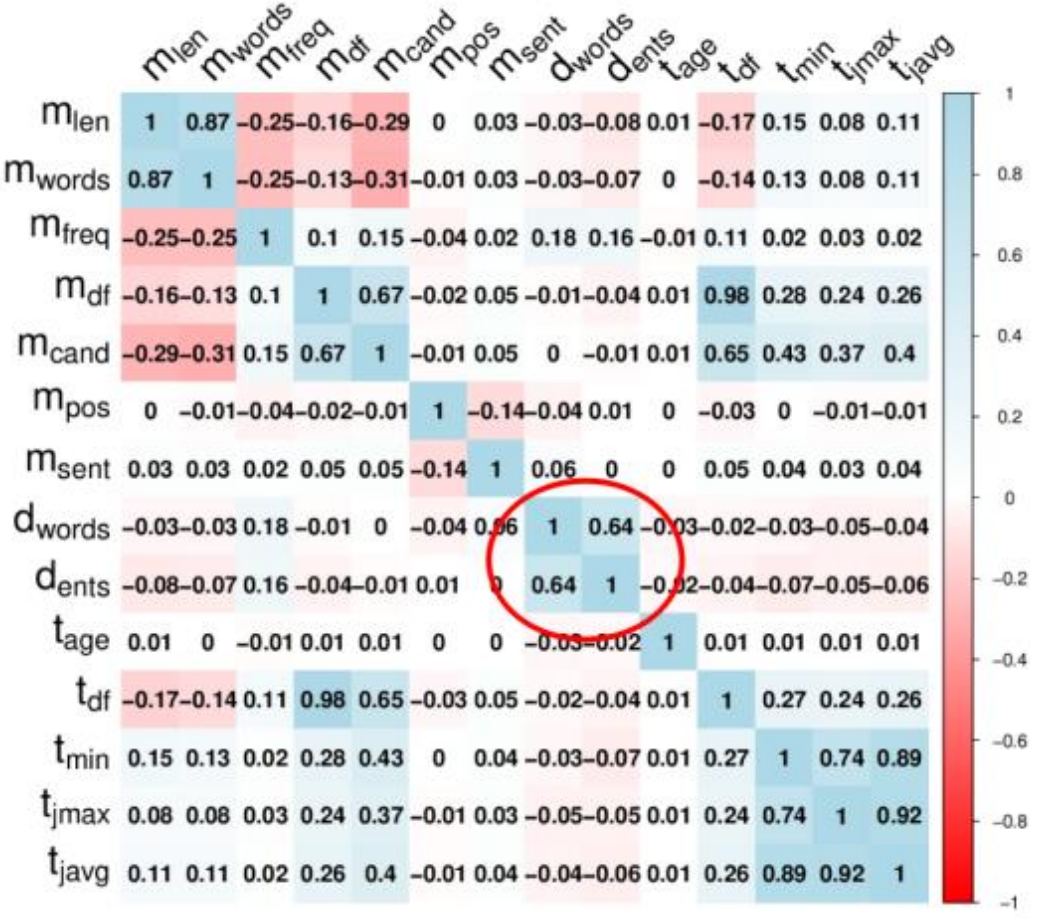
Feature Analysis

Correlation among features – **Pearson's ρ**



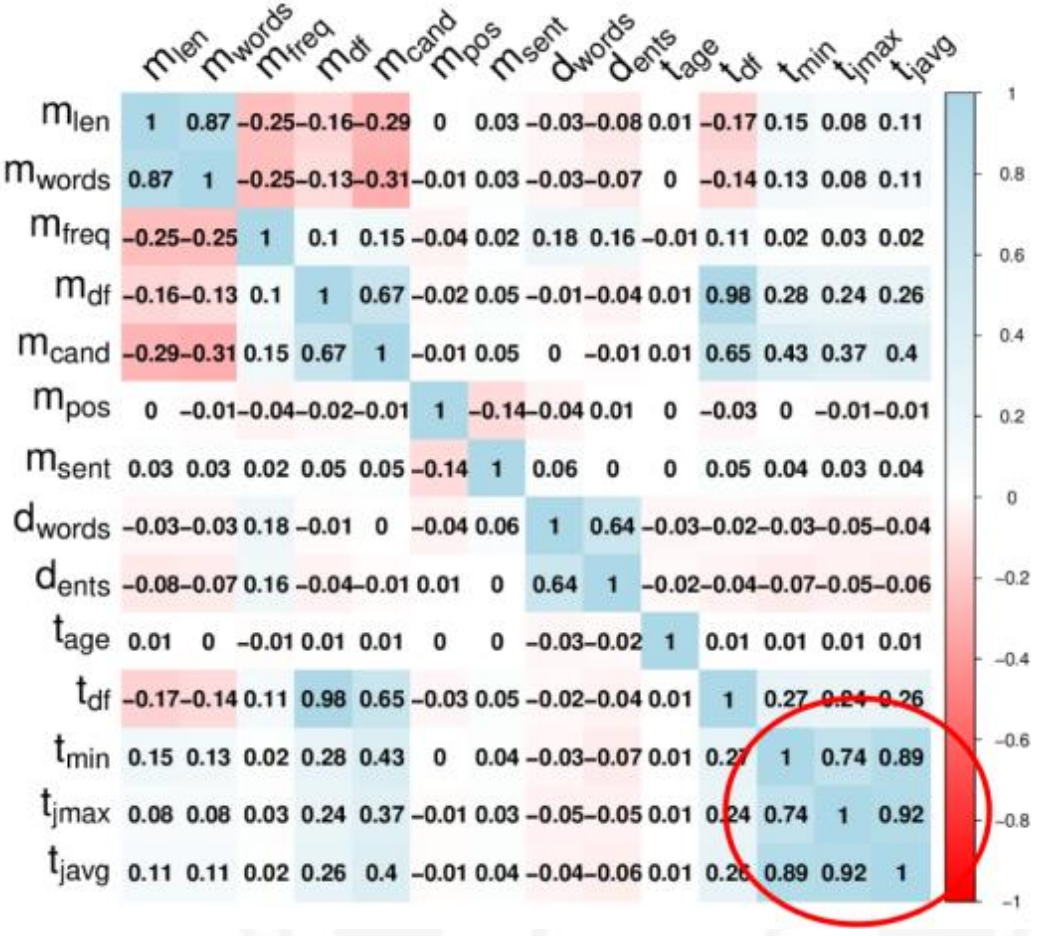
Feature Analysis

Correlation among features – **Pearson's ρ**



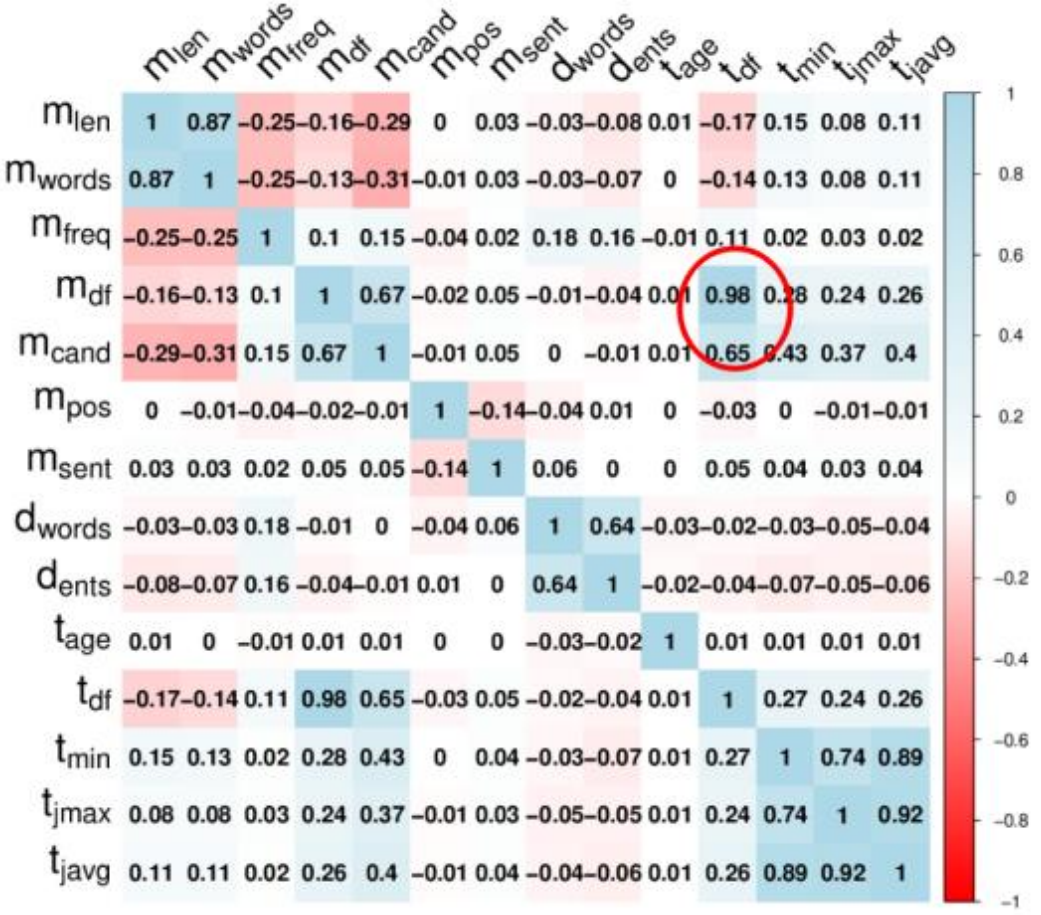
Feature Analysis

Correlation among features – **Pearson's ρ**



Feature Analysis

Correlation among features – **Pearson's ρ**



Impact on Entity Linking

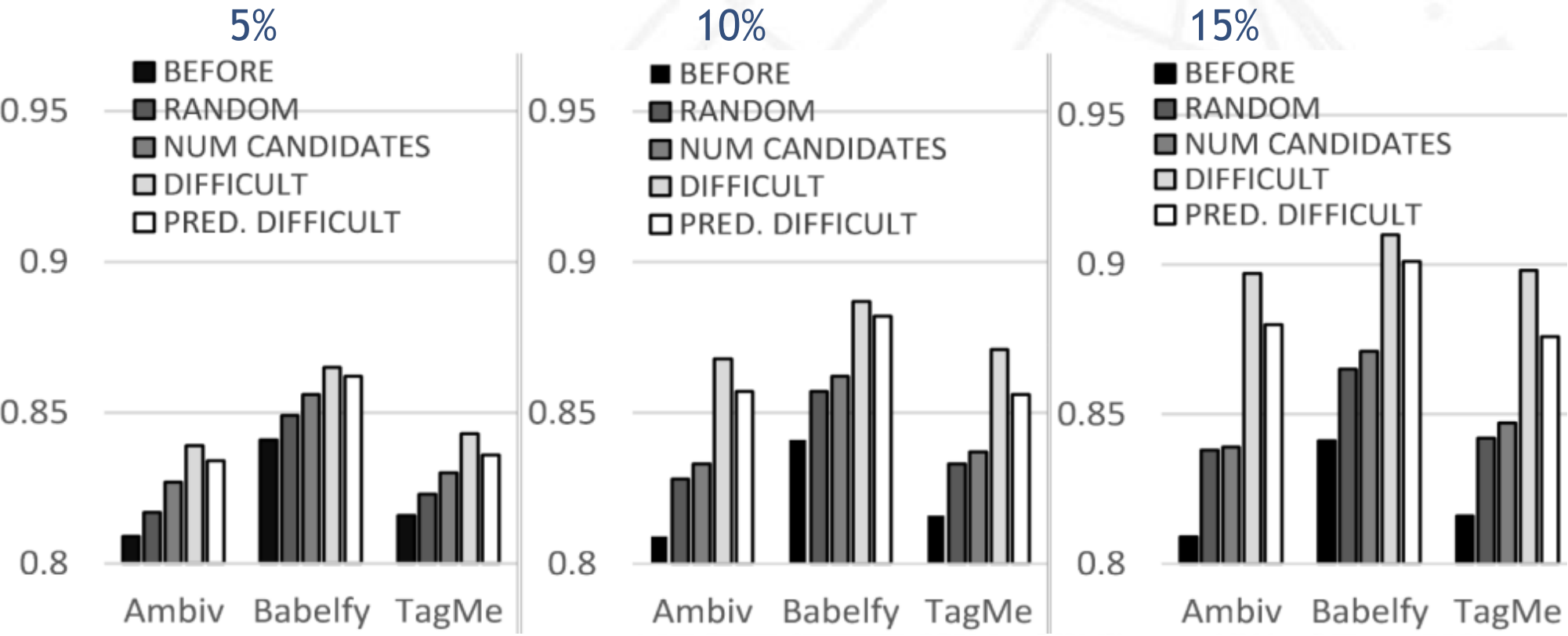
- Assessing overall performance of a semi-automated EL pipeline
- Simulation: Human annotators are guided by our approach to complement entity links with manual annotation of **HARD** cases

- **CONLL-TestB**

- Select *N* mentions for manual annotations
  - **BEFORE**
  - **RANDOM**
  - **CANDIDATES**
  - **DIFFICULT**
  - **PRED. DIFFICULT**

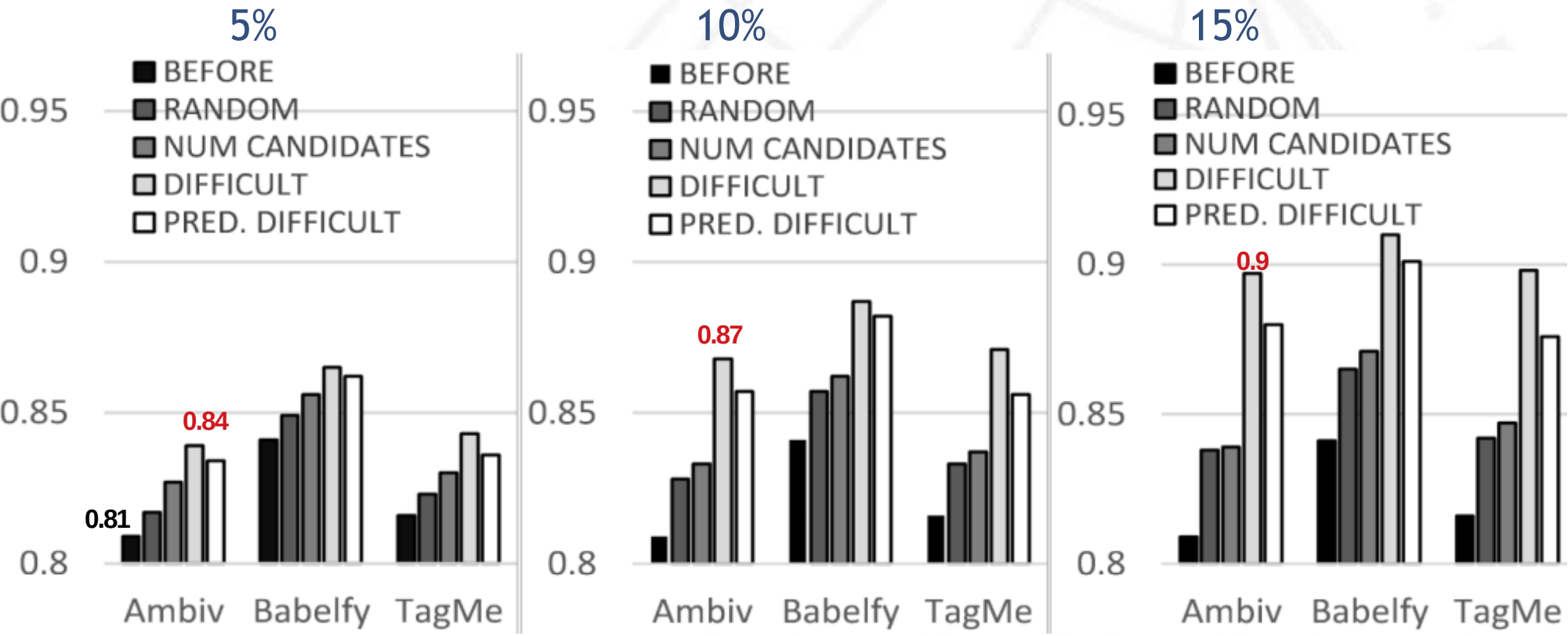
Impact on Entity Linking

- Effect of human feedback on the accuracy of semi-automated EL systems for different proportion of human judgments



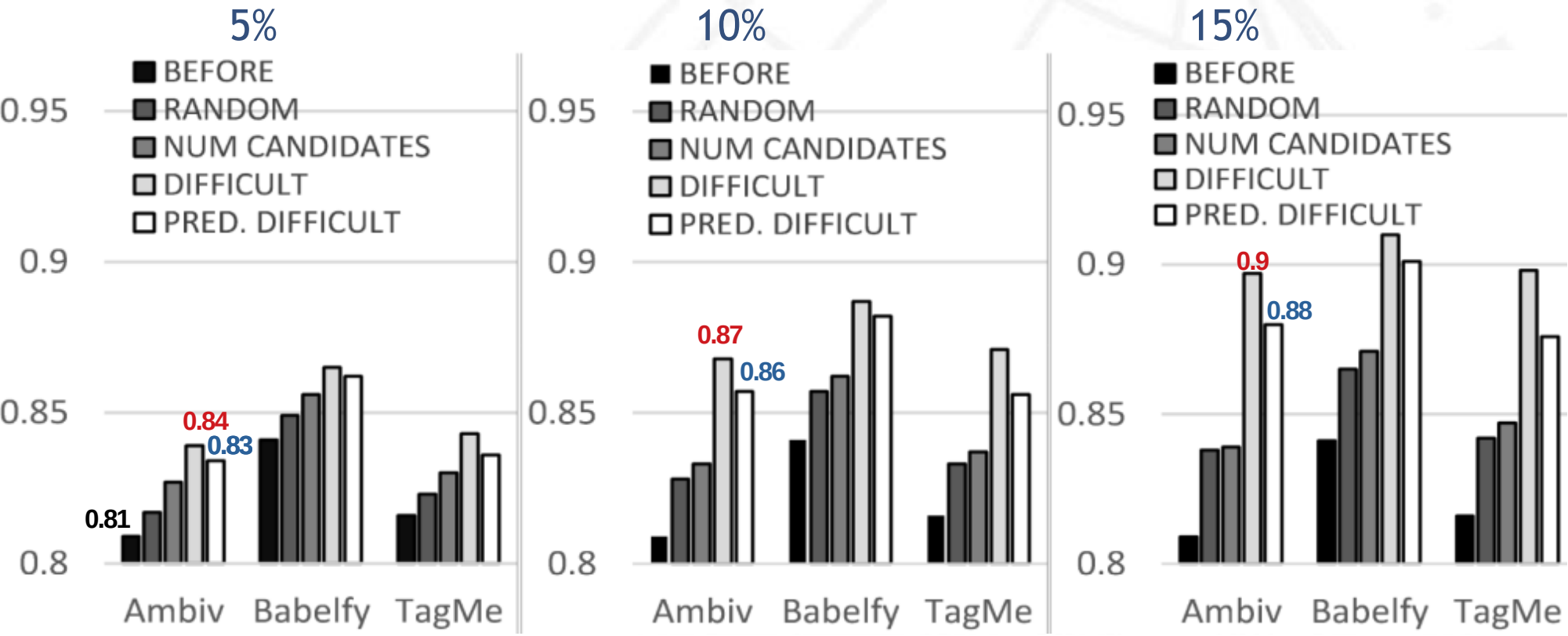
Impact on Entity Linking

- Effect of human feedback on the accuracy of semi-automated EL systems for different proportion of human judgments



Impact on Entity Linking

- Effect of human feedback on the accuracy of semi-automated EL systems for different proportion of human judgments



Conclusions and Future work

Conclusions

- Novel problem of detecting and understanding EL difficulty
  - Ambiverse accuracy is increased *from 0.81 to 0.87* when **10%** of the recognized mentions labelled as **HARD** are manually judged
- Introduced a set of features which can be used within a distantly supervised model for predicting difficult to link mentions
  - Difficulty labels can be predicted with $P > 0.83$
- Prediction can be used to detect latent characteristics that affect EL performance
  - *NYT corpus* – The mention's position characterizes many **HARD** cases

Future Works

- Investigate more features
  - Document fluency
  - Lexical diversity
  - Mention's semantic evolution
- Investigate effectiveness of other oversampling methods
  - SMOTE
- Investigate cost-sensitive classifiers
  - Focus on increasing minority class' performance

Thank you !

Questions?