Applications of Text Mining & NLP

Javier Garcia-Bernardo

Summary

Text -> Numeric:

- TF; TF-IDF; Word embeddings

Similarity (often cosine similarity)

Clustering

- TF-IDF + Dimensionality reduction (e.g. TruncatedSVD) + Clustering (e.g. K-Means)
- Word embeddings + Clustering (e.g. K-Means)
- Document-Term Matrix (CountVectorizer) + LDA

Classification

- TF-IDF (+ Dimensionality reduction) + Classifier (e.g. LogisticRegression)
- Word embeddings + Classifier (e.g. LogisticRegression)
- Neural Networks (Feed-forward, RNN, LSTM, CNN, Transformers)

Tomorrow: **Sentiment analysis**

A collection of text mining applications

Similarity

- Find authors of an anonymous book
- Analyze experiments on information diffusion
- Find duplicates

Clustering

- Targeted advertisement or learning or helper
- Recommendation systems (e.g. similar books)
- Clustering stories (clustering fiction works, people's diagnoses, misinformation)
- Track evolution of topics in discourse
- Anomaly detection

Classification/Regression

- Targeted advertisement
- Hate speech classification (similar: spam, fake news, emergencies)
- Tracking emotions (some emotions are highly correlated with YouGov surveys)
- Predict student performance, probability of re-hospitalization, fraud (insurance or research)
- Classifying reports (e.g. hospital discharges, urgent issues)
- Predict stock market returns

Applications in Social Media (Hate Speech in Twitter)

Hateful Symbols or Hateful People?
Predictive Features for Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Using Convolutional Neural Networks to Classify Hate-Speech

Björn Gambäck and **Utpal Kumar Sikdar**
Department of Computer Science
Norwegian University of Science and Technology
NO–7491 Trondheim, Norway
gamback@ntnu.no utpal.sikdar@gmail.com

Chapter 3
Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection

Zeerak Waseem, James Thorne and Joachim Bingel

A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media

Marzieh Mozafari(✉), Reza Farahbakhsh, and Noël Crespi

Waseem and Hovy, 2016

Data: Annotation of 16k tweets based on Gender studies and CRT

Method: TD-IDF using character {uni, bi, tri}-grams.

Why did they use characters instead of words?

Preprocessing: Removing stop words (except "not"), usernames and punctuation

Classifier: Logistic Regression

Results:

System setup	Precision	Recall	F_1-score
Logistic Regression with character n-grams	0.7287	**0.7775**	0.7389

A tweet is offensive if it

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

Gamback and Sikdar, 2017

Data: Waseem&Hovy (2016)

Method: CNN using word embeddings and character n-grams

Preprocessing: None

Classifier: Softmax

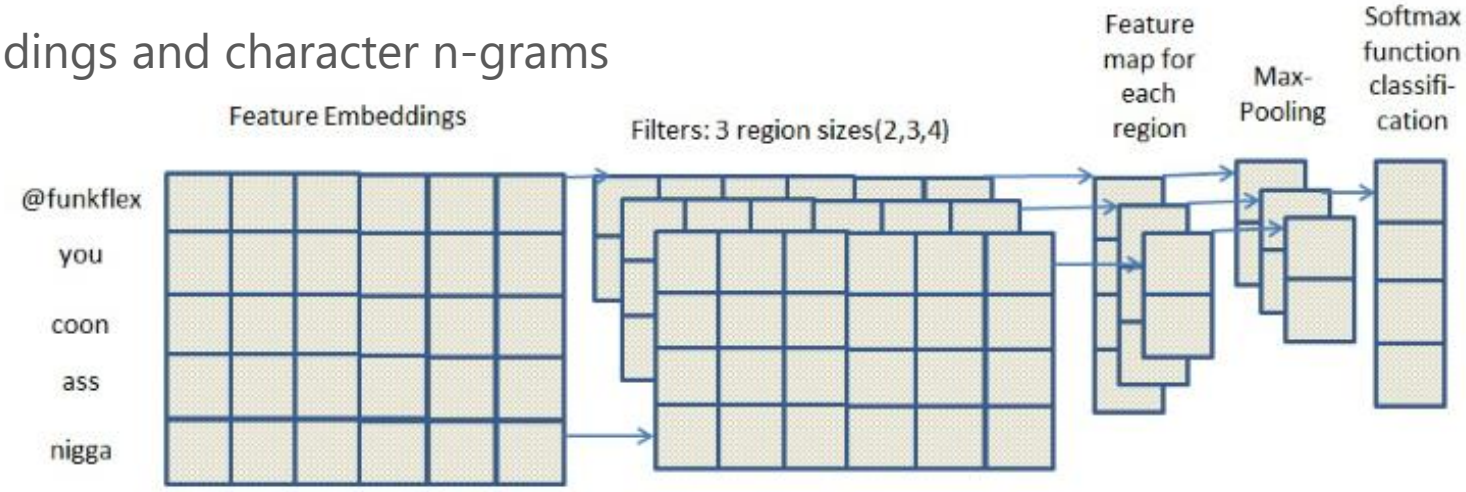


Figure 1: Hate-speech classifier

Results:

	System setup	Precision	Recall	F_1-score
CNN	Random vectors	**0.8668**	0.6726	0.7563
CNN	word2vec	0.8566	0.7214	**0.7829**
CNN	Character n-grams	0.8557	0.7011	0.7695
CNN	word2vec + character n-grams	0.8661	0.7042	0.7738
	Logistic Regression with character n-grams (Waseem and Hovy, 2016)	0.7287	**0.7775**	0.7389

Table 2: System performance (10-fold cross-validated)

Waseem, Thorne and Bingel (2018)

Data: Waseem & Hovy (2016), 25k annotated tweets (Davidson et al, 2017; Twitter user guidelines)

Best method: Multi-task training. BoW words (5000), bigrams (5000) and character bi/tri-gram (5000)

Feed-forward neural network with 2 hidden layesr. Softmax output

Preprocessing: Removing usernames, links and punctuation

Classifier: Softmax

Regularization: Dropout (20%)

Results:

Method	Datasets	Precision(%)	Recall(%)	F1-Score(%)
Waseem and Hovy [22]	Waseem	72.87	77.75	73.89
Davidson et al. [3]	Davidson	91	90	90
Waseem et al. [23]	Waseem	-	-	80
	Davidson	-	-	89

Mozafari, Farahbakhsh and Crespi (2019)

Data: Waseem & Hovy (2016), 25k annotated tweets (Davidson et al, 2017; Twitter user guidelines)

Best method: BERT + CNN

Each layer of the transformer gives an output → CNN

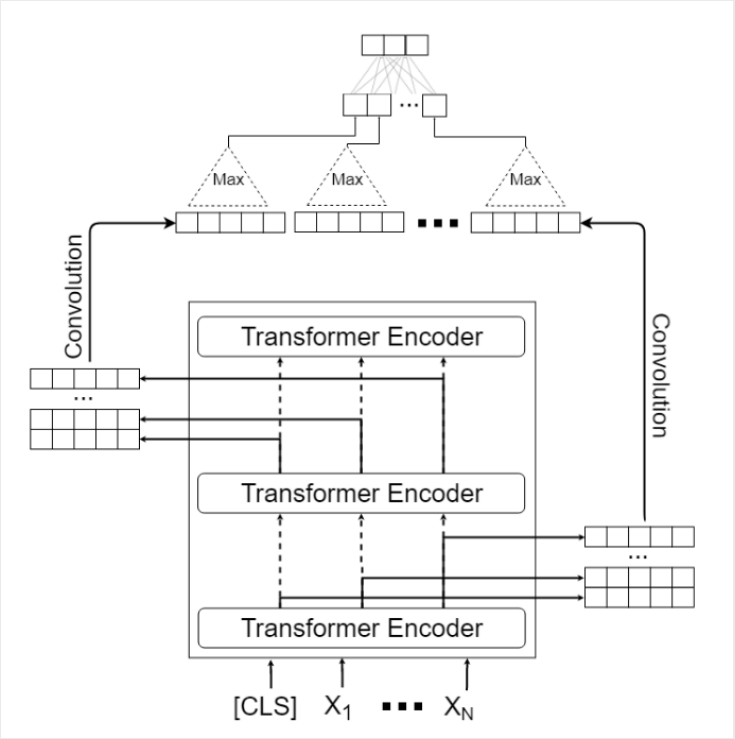
Preprocessing: Replacing usernames, elongated words, hashtags; remove punctuation

Classifier: Softmax

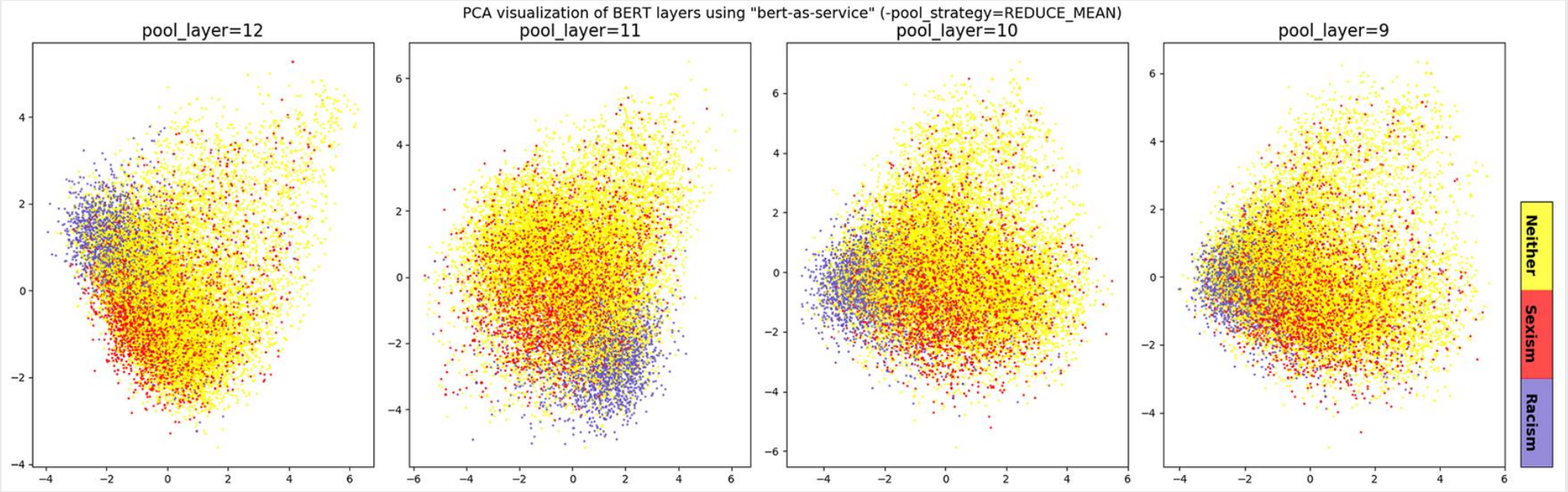
Regularization: Dropout (10%)

Results:

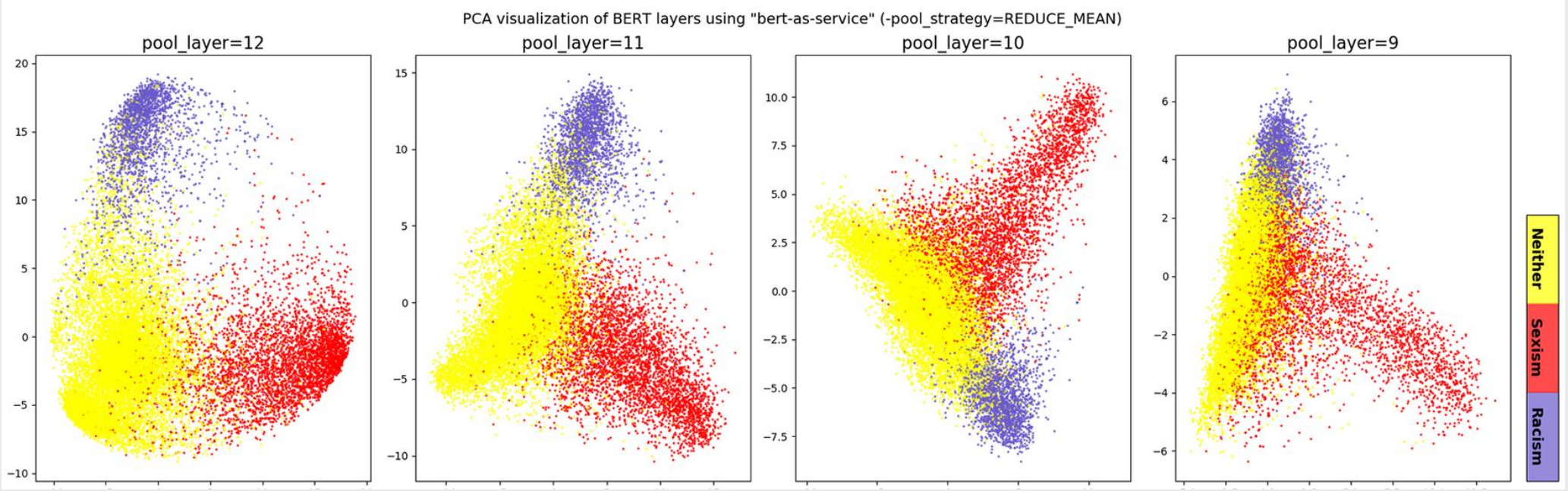
Method	Datasets	Precision(%)	Recall(%)	F1-Score(%)
Waseem and Hovy [22]	Waseem	72.87	77.75	73.89
Davidson et al. [3]	Davidson	91	90	90
Waseem et al. [23]	Waseem	-	-	80
	Davidson	-	-	89
$BERT_{base}$	Waseem	81	81	81
	Davidson	91	91	91
$BERT_{base}$ + Nonlinear Layers	Waseem	73	85	76
	Davidson	76	78	77
$BERT_{base}$ + LSTM	Waseem	87	86	86
	Davidson	91	92	92
$BERT_{base}$ + CNN	Waseem	**89**	**87**	**88**
	Davidson	**92**	**92**	**92**



Base BERT



Fine-tuned BERT



Application of Text Clustering in Media

RESEARCH ARTICLE

Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter

Philipp Wicke[1]*, **Marianna M. Bolognesi**[2]

Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach

Tobias Heidenreich ✉, Fabienne Lind, Jakob-Moritz Eberl, Hajo G Boomgaarden

Journal of Refugee Studies, Volume 32, Issue Special_Issue_1, December 2019, Pages i172–i182, https://doi.org/10.1093/jrs/fez025

Published: 27 December 2019 **Article history** ▾

COVID pandemic (Wicke & Bolognesi 2020)

Question: What is the framing of the COVID pandemic? Framing of WAR (fight, combat, battle), STORM (wave, storm, cloud), MONSTER (evil, horror, killer) or TSUNAMI (wave, tragedy, catastrophe).

Data: Twitter around #Covid-19 (80 hashtags)

Method: LDA (4 and 16 topics) + correlation of topics with frames

Preprocessing: Remove stop words, remove covid, remove tokes with less than 3 characters

Validity assessed: Qualitatively + external database.

Results:

Refugees crisis (Heidenreich, Lind, Eberl & Boomgaarden, 2019)

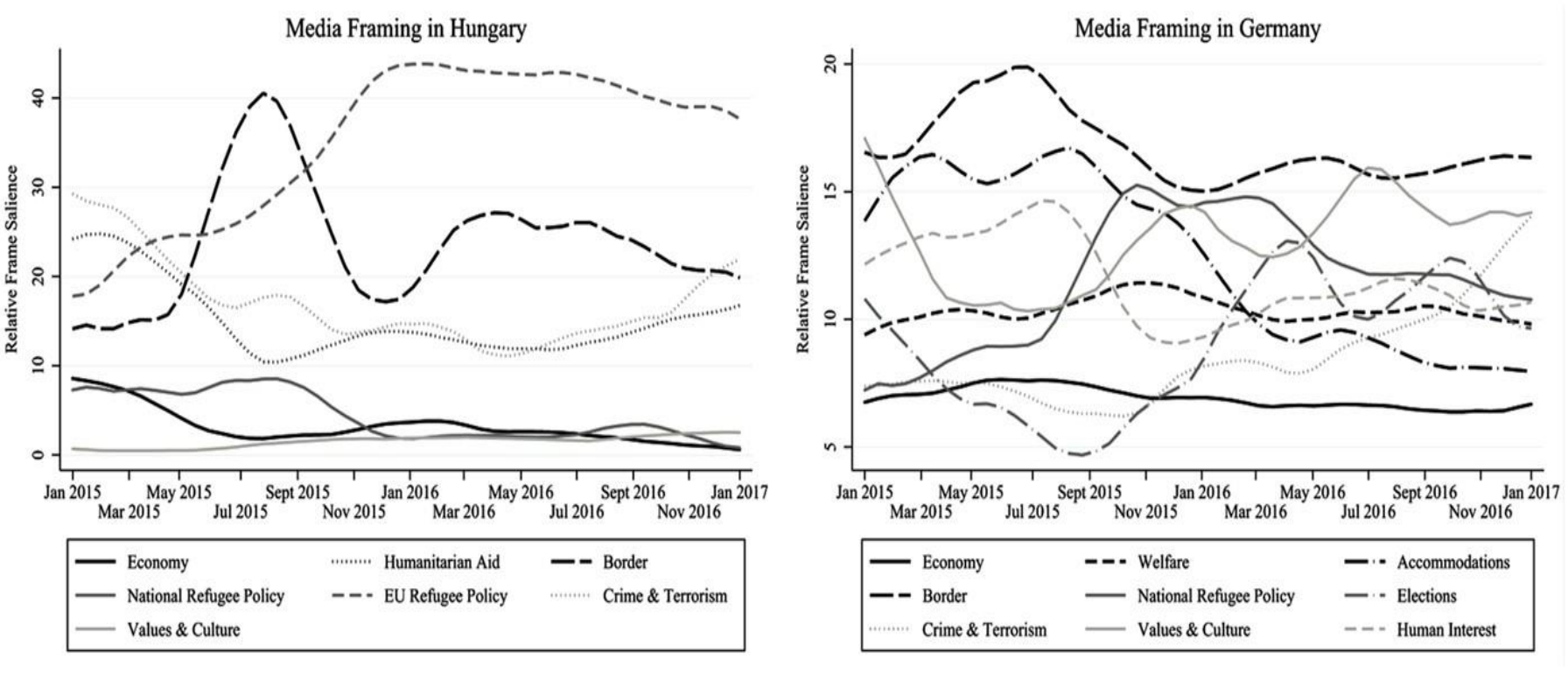
Data: 130k articles from 24 news outlets

Method: LDA (10 topics per country) + manual labeling.

Preprocessing: Unclear

Validation: Semantic validity (are the topics distinctive) + Randomly reading three articles per topic/country + predictive validity (are important events such as elections reflected)

Results:



Similarity between words

Applications in Information Diffusion

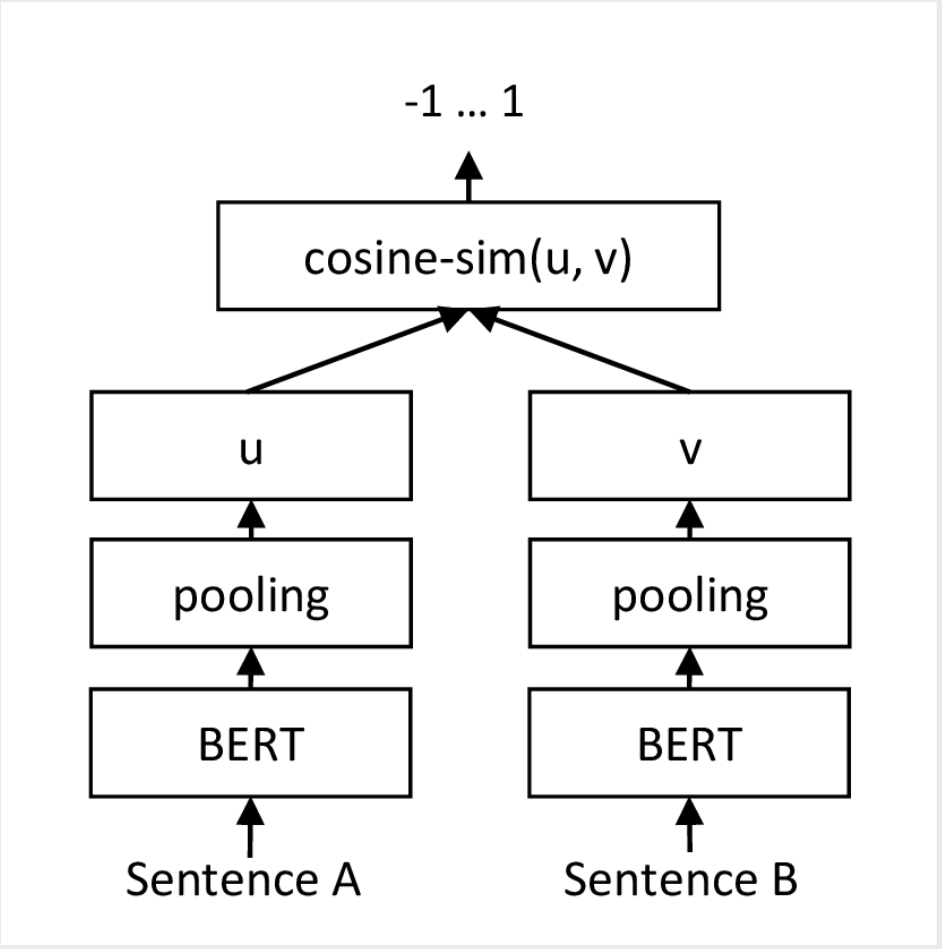
How to measure similarity between words?

Option 1: Cosine similarity TF-IDF representations

Option 2: Average word embeddings, cosine similarity. Works only for small sentences.

Option 3: Train BERT, pass two sentences (but creates one embedding for the combination)

Option 4: Siamese BERT-Networks: Pre-trained only on books + Wikipedia (via BERT) and on NLI data (sentence pairs)



Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Sentence similarity

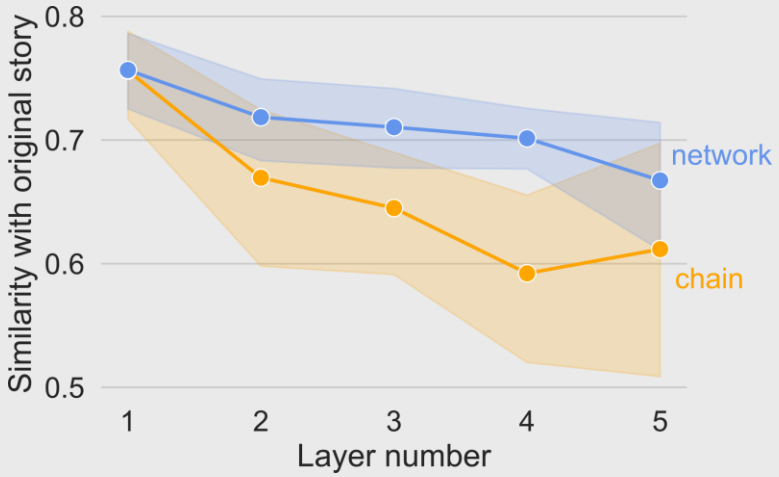
Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

Application: Analyzing experiment data

Aloric, Garcia-Bernardo, Krafft, Hardy, Morgan, Neu, Santoro (202x)

Question: Understand effect of network structure on information diffusal

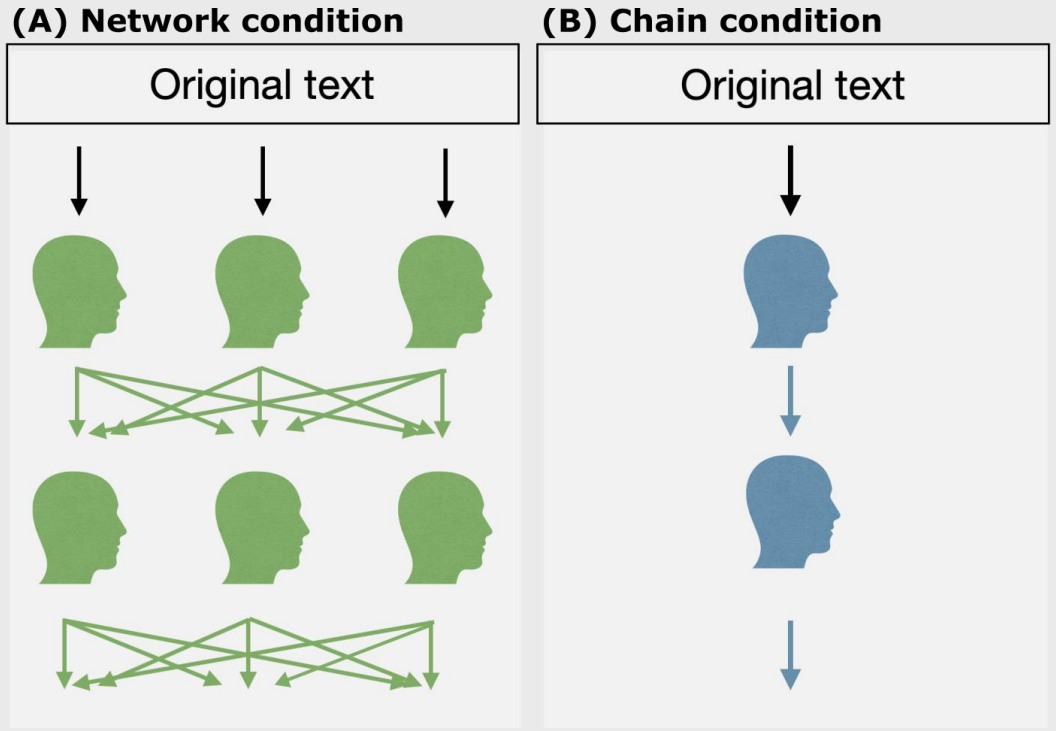
Hypothesis 1: Network helps preserve more information



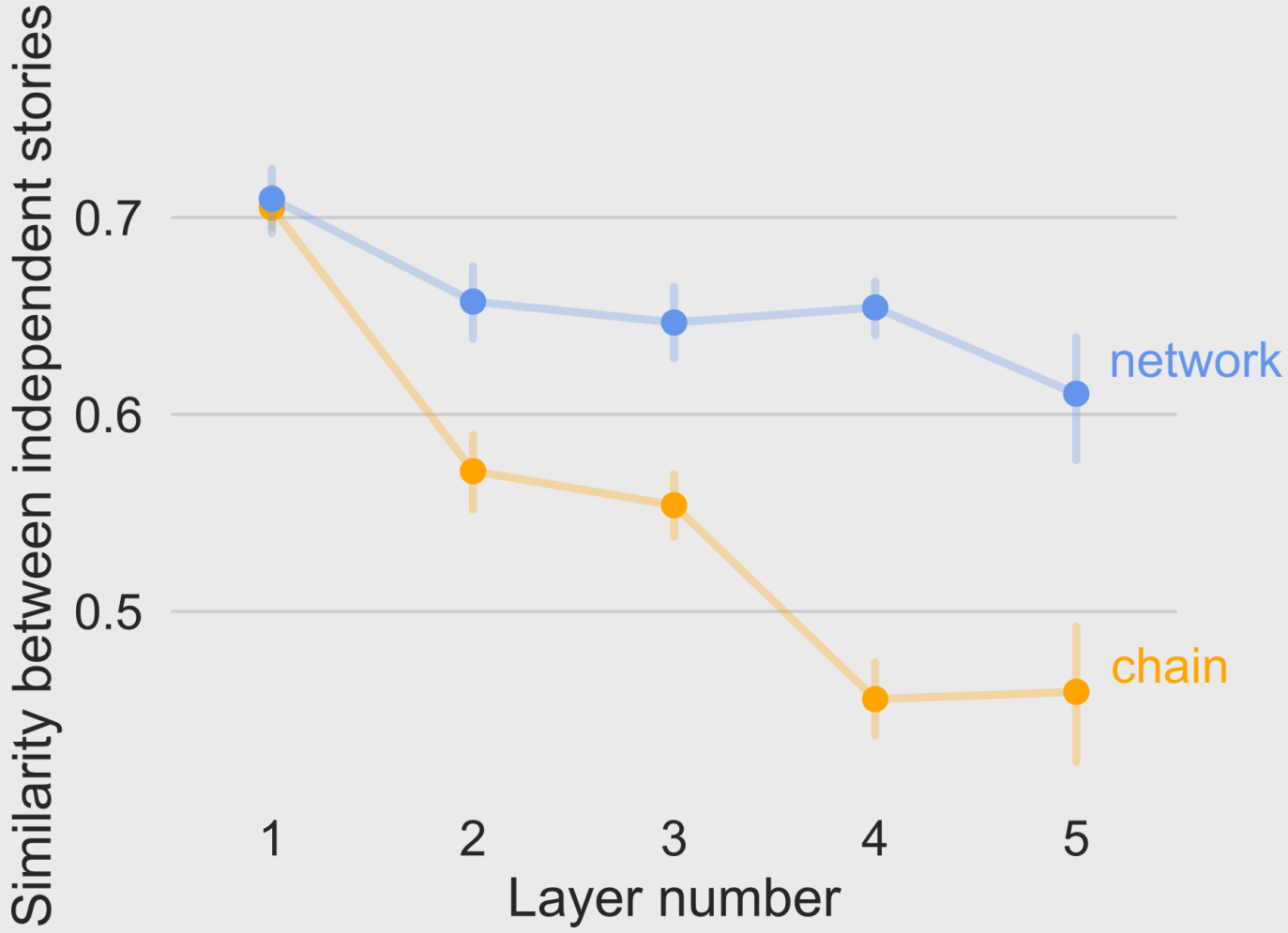
Hypothesis 2: Network structure helps preserve "the essence" of the story

If true: Independent replicates should be more correlated

Data/experimental design:



Between-replicates similarity



Applications in Humanities

The emotional arcs of stories are dominated by six basic shapes

Andrew J Reagan ✉, Lewis Mitchell, Dilan Kiley, Christopher M Danforth & Peter Sheridan Dodds

EPJ Data Science **5**, Article number: 31 (2016) | Cite this article

‘There is no reason why the simple shapes of stories can’t be fed into computers, they are beautiful shapes.’
- Kurt Vonnegurt

Start with all 51,250 Project Guterberg eBooks

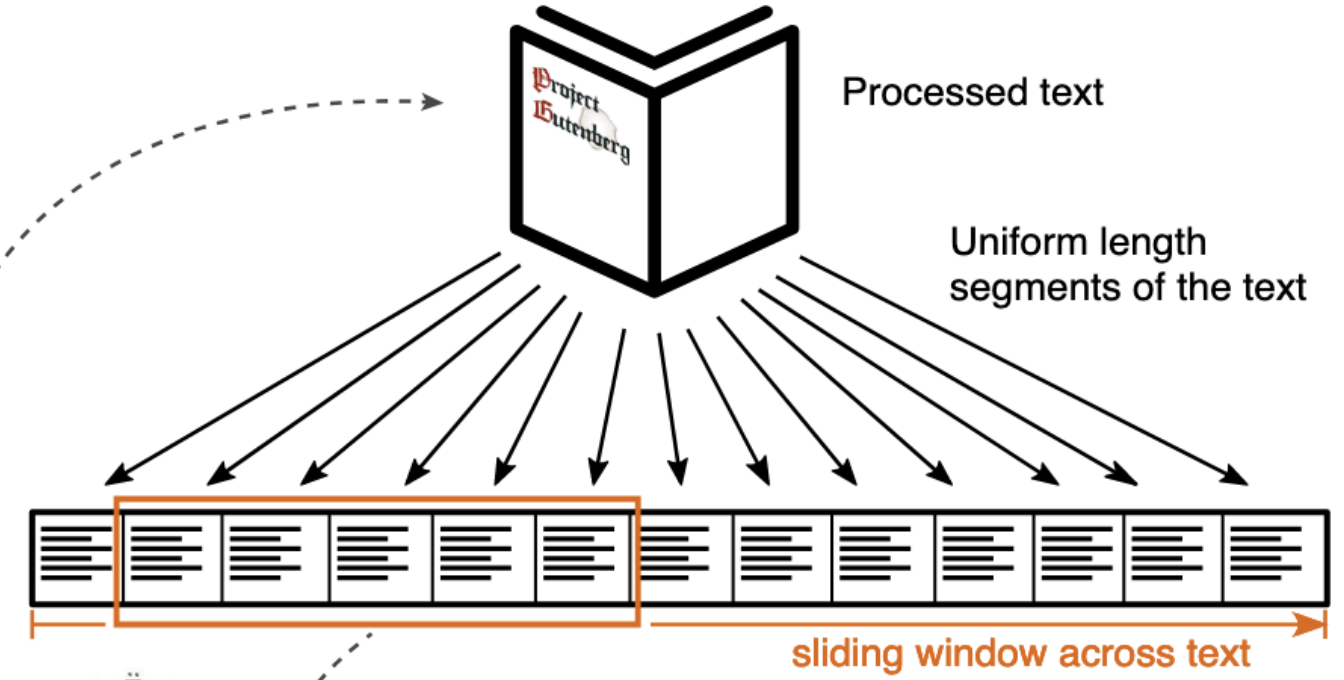
Select for Library of Congress Class ``Fiction''

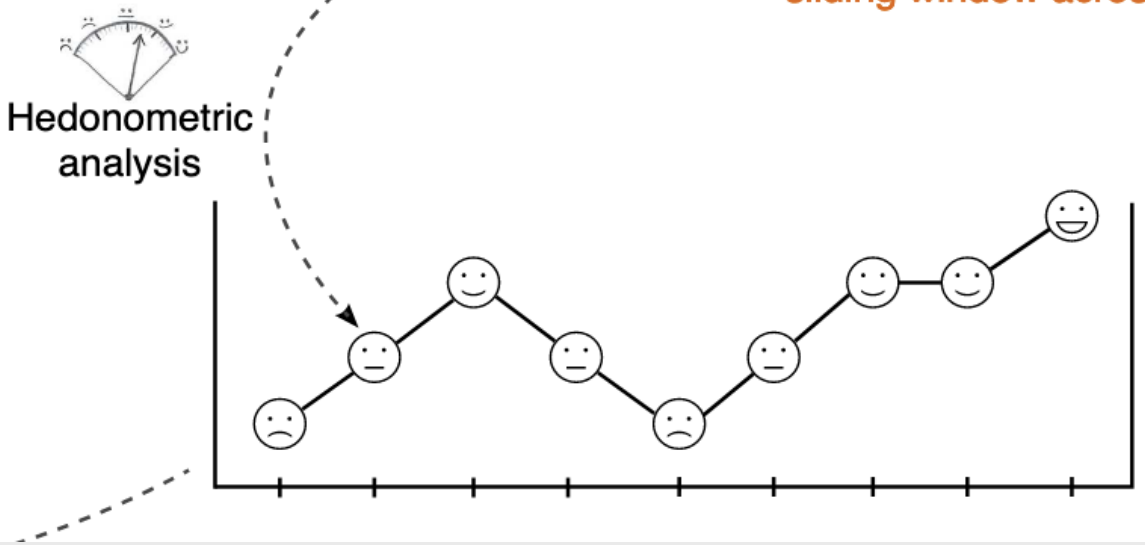
18,561 fiction eBooks

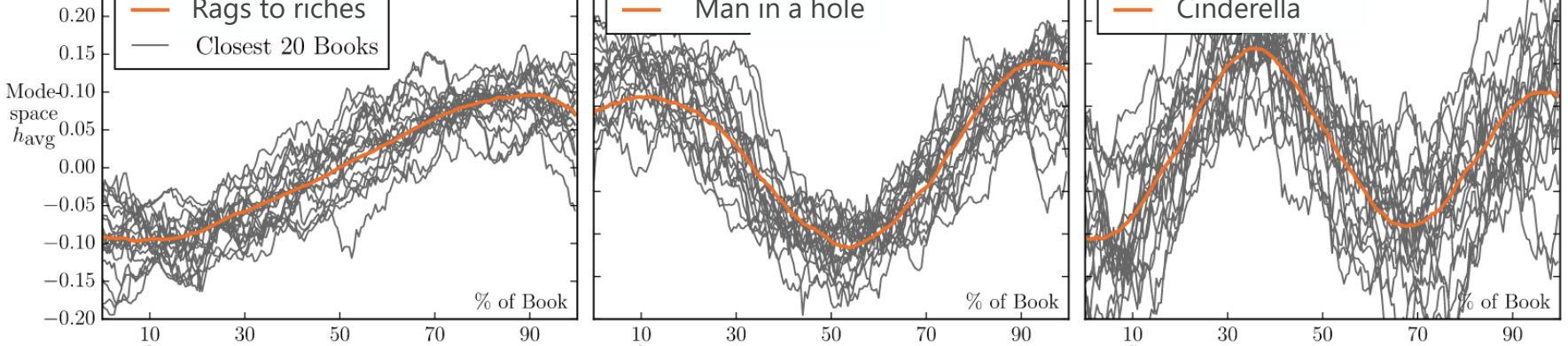
Filter for length, number of downloads, English, duplicates, and collec- tions

1,385 English fiction eBooks

Remove header/footer, correct file encodings, split into chapters, sentences, and tokenize words







Top Stories:

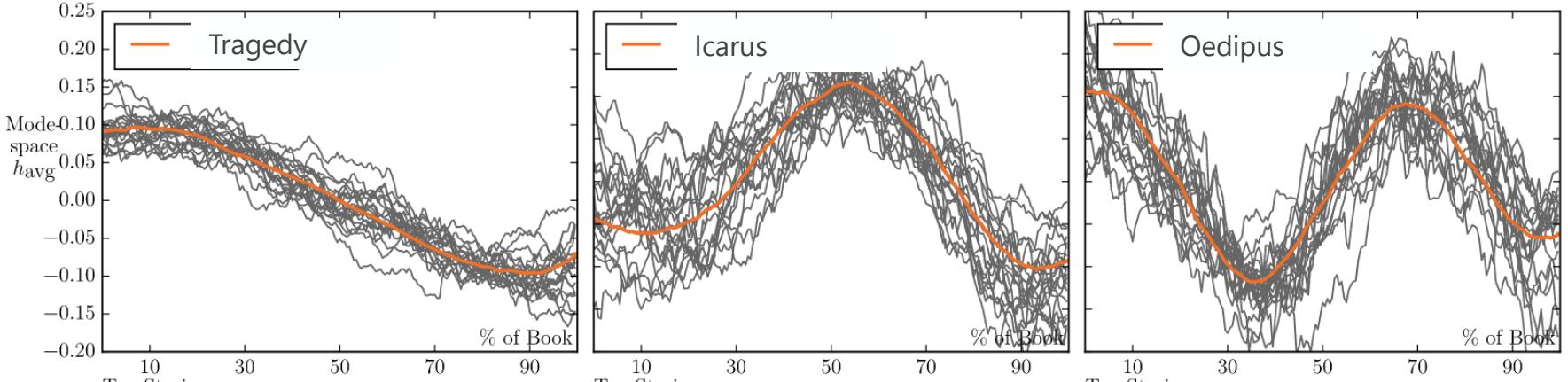
1: The Winter's Tale (1539, 73)
http://hedonometer.org/books/v3/1539/
2: Oscar Wilde, Art and Morality: A... (33689, 88)
http://hedonometer.org/books/v3/33689/
3: The Terror: A Mystery (35617, 61)
http://hedonometer.org/books/v3/35617/
4: The Pilgrim's Progress in Words ... (7088, 55)
http://hedonometer.org/books/v3/7088/
5: The Road to Oz (26624, 68)
http://hedonometer.org/books/v3/26624/

Top Stories:

1: The Magic of Oz (419, 186)
http://hedonometer.org/books/v3/419/
2: Children of the Frost (10736, 82)
http://hedonometer.org/books/v3/10736/
3: Tamburlaine the Great — Part 1 (1094, 474)
http://hedonometer.org/books/v3/1094/
4: The Life and Adventures of Santa... (520, 76)
http://hedonometer.org/books/v3/520/
5: Justice (2911, 50)
http://hedonometer.org/books/v3/2911/

Top Stories:

1: The Mystery of the Hasty Arrow (17763, 93)
http://hedonometer.org/books/v3/17763/
2: Through the Magic Door (5317, 81)
http://hedonometer.org/books/v3/5317/
3: After London; Or, Wild England (13944, 146)
http://hedonometer.org/books/v3/13944/
4: The Shadow of the Rope (12590, 75)
http://hedonometer.org/books/v3/12590/
5: That Affair at Elizabeth (35247, 62)
http://hedonometer.org/books/v3/35247/



Top Stories:

1: Lady Susan (946, 894)
http://hedonometer.org/books/v3/946/
2: Warlord of Kor (17958, 70)
http://hedonometer.org/books/v3/17958/
3: The House of the Vampire (17144, 188)
http://hedonometer.org/books/v3/17144/
4: Tom Sawyer, Detective (93, 112)
http://hedonometer.org/books/v3/93/
5: The Island of Doctor Moreau (159, 1083)
http://hedonometer.org/books/v3/159/

Top Stories:

1: Shadowings (34215, 63)
http://hedonometer.org/books/v3/34215/
2: Battle-Pieces and Aspects of the... (12384, 194)
http://hedonometer.org/books/v3/12384/
3: The Slayer of Souls (36281, 63)
http://hedonometer.org/books/v3/36281/
4: The Bobbsey Twins : Or, Merry Day... (17412, 69)
http://hedonometer.org/books/v3/17412/
5: Allan's Wife (2727, 128)
http://hedonometer.org/books/v3/2727/

Top Stories:

1: This World Is Taboo (18172, 64)
http://hedonometer.org/books/v3/18172/
2: Old Indian Days (339, 139)
http://hedonometer.org/books/v3/339/
3: The Evil Guest (10377, 93)
http://hedonometer.org/books/v3/10377/
4: Pariah Planet (29448, 96)
http://hedonometer.org/books/v3/29448/
5: The Wind in the Willows (289, 1475)
http://hedonometer.org/books/v3/289/

Consideration of the emotional arc for a given story is important for the success of that story

Applications in Health: Automating coding

ICD-10 Coding of Spanish Electronic Discharge Summaries: An Extreme Classification Problem

Publisher: IEEE Cite This PDF

Mario Almagro ; Raquel Martínez Unanue ; Víctor Fresno ; Soto Montalvo **All Authors**

Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks

Arjan Sammani, Ayoub Bagheri, Peter G. M. van der Heijden, Anneline S. J. M. te Riele, Annette F. Baas, C. A. J. Oosters, Daniel Oberski & Folkert W. Asselbergs

npj Digital Medicine **4**, Article number: 37 (2021) | Cite this article

Almagro, Martínez-Unanue, Fresno, Moltalvo, 2020

Goal: Suggest a list of the 10 most probable ICD-10 codes (diseases, abnormal findings, causes of injury…) to experts

Data: 7k discharged reports, with 7k ICD-10 codes. Cardinality= 10.

Method: Different methods

Preprocessing: Remove sentences withoug technical terms (using tagging software), removal accents, punctuation, stemming.

Results:

Method	*P*@10
Baseline	14.59
SVMs	37.06
MLPs	35.28
AdaBoost	36.36
GBoost	**40.88**
KLD	16.52
Document-Similarity	29.37
LSTM	15.08
XML-CNN	24.99
FastXML	29.87
SLEEC	27.00
Dependency-LDA	31.96
Voting	**46.75**
Regression	41.73

Bagheri, Sammani, van der Heijden, Asselbergs, Oberski, 2020

Question: The proposal is conceived to be applied in a real system, suggesting a list of the 10 most probable codes to experts

Data: 6k discharged reports, with 1k ICD-10 (diseases, abnormal findings, causes of injury…). Cardinality=5

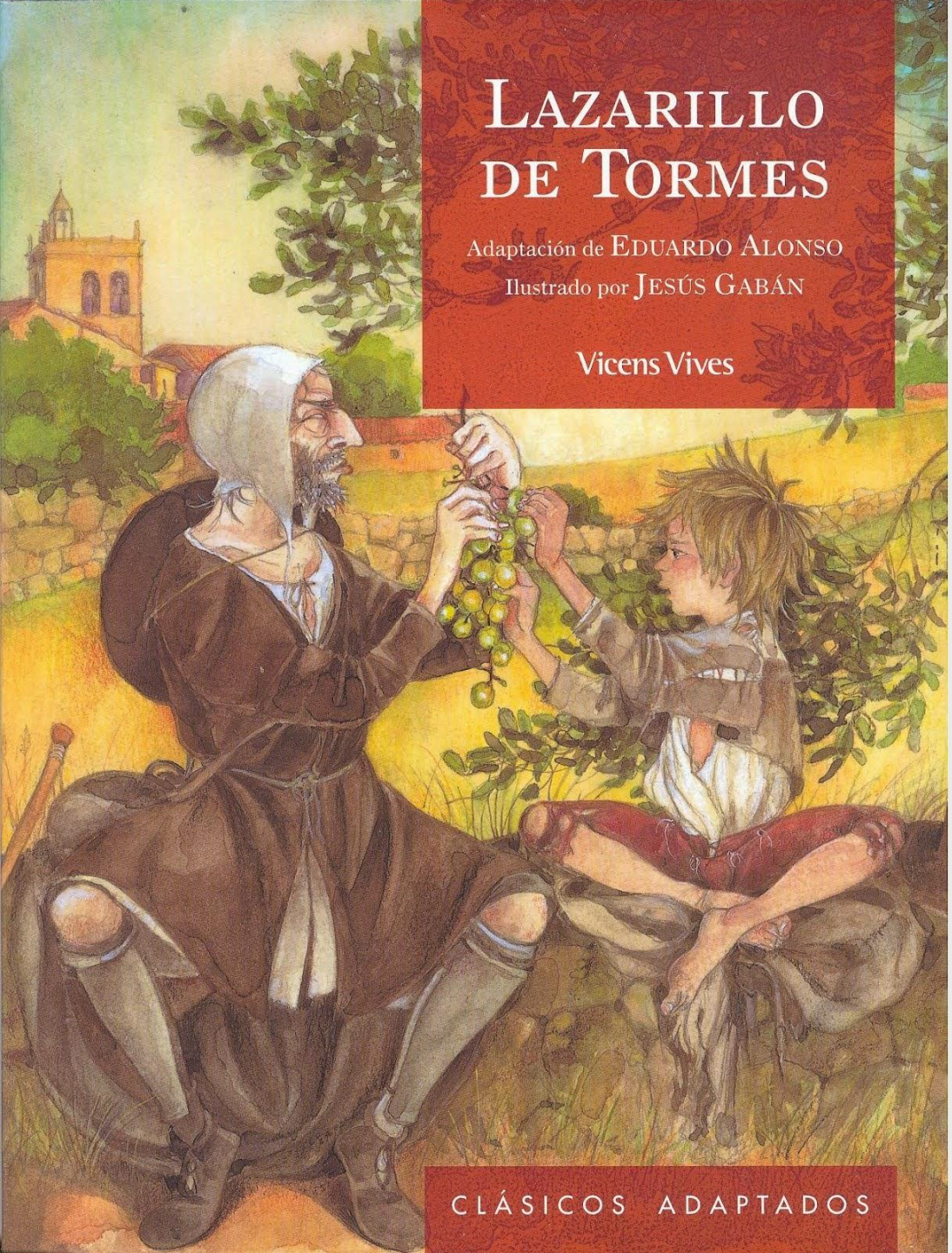
Method: Different methods

Preprocessing: removed small labels, trimmed whitespaces, numbers and converted all characters to lowercase

Results:

	ICD chapters		**Rolled-up ICD codes**	
	Accuracy	F1	Accuracy	F1
BOW SVM (baseline)	**62.3**	**74.3**	11.6	20.2
Average word embeddings (SVM)	60.4	72.6	12.5	**25.8**
CNN(1conv)	38.1	46.3	09.0	16.1
CNN(2conv)	42.2	49.0	12.4	19.1
LSTM	53.4	59.6	11.7	18.8
BiLSTM	55.0	70.1	13.7	23.2
HA-GRU	56.8	71.3	**15.9**	24.3

Detecting unique patterns in text



Year	Author
1605	Juan de Ortega†
1607	Diego Hurtado de Mendoza†
1608	Diego Hurtado de Mendoza
1624	Juan de Ortega
	Diego Hurtado de Mendoza
1867	Sebastián de Horozco†
1873	Diego Hurtado de Mendoza
1888	Diego Hurtado de Mendoza
	Juan de Valdés†
1901	Lope de Rueda†
1914	Juan de Valdés
	Lope de Rueda
	Sebastián de Horozco[75]
1915	Sebastián de Horozco
1943	Diego Hurtado de Mendoza
	Diego Hurtado de Mendoza
1954	Juan de Ortega
1955	Pedro de Rhúa†
1957	Sebastián de Horozco
1959	Juan de Valdés
	Juan de Valdés
1960	Juan de Valdés
1961	Diego Hurtado de Mendoza
1963	Diego Hurtado de Mendoza
1964	Hernán Núñez de Toledo†
	Lope de Rueda
1966	Juan de Ortega
1969	Diego Hurtado de Mendoza
1970	Diego Hurtado de Mendoza
1973	Sebastián de Horozco
1976	Alfonso de Valdés†
1978	Sebastián de Horozco
1980	Lope de Rueda
	Sebastián de Horozco
1987	Lope de Rueda
	Sebastián de Horozco
	Hernán Núñez de Toled
1988	Juan de Ortega
1992	Juan de Valdés
2002	Alfonso de Valdés
	Alfonso de Valdés
	Juan de Ortega
2003	Lope de Rueda
	Alfonso de Valdés
	Alfonso de Valdés
	Francisco Cervantes de Salazar†
	Alfonso de Valdés
	Alfonso de Valdés
2004	Alfonso de Valdés
	Alfonso de Valdés
	Alfonso de Valdés
2006	Alfonso de Valdés
	Alfonso de Valdés
	Alfonso de Valdés
	Lope de Rueda
	Juan Luis Vives†
2007	Alfonso de Valdés
2008	Pedro de Rhúa
	Francisco Cervantes de Salazar
	Juan Arce de Otálora†
2010	Diego Hurtado de Mendoza
	Diego Hurtado de Mendoza
	Diego Hurtado de Mendoza
	Alfonso de Valdés
	Alfonso de Valdés
	Diego Hurtado de Mendoza
	Diego Hurtado de Mendoza
	Juan Arce de Otálora
	Juan Arce de Otálora
	Juan de Pineda†
2011	Juan Arce de Otálora
	Diego Hurtado de Mendoza
	Diego Hurtado de Mendoza
2012	Juan Luis Vives
2014	Juan Luis Vives
	Diego Hurtado de Mendoza
	Juan Arce de Otálora

The Life of *Lazarillo de Tormes* and of His Machine Learning Adversities

Non-traditional authorship attribution techniques in the context of the *Lazarillo*

Javier de la Rosa & Juan Luis Suárez
The University of Western Ontario, London, Canadá

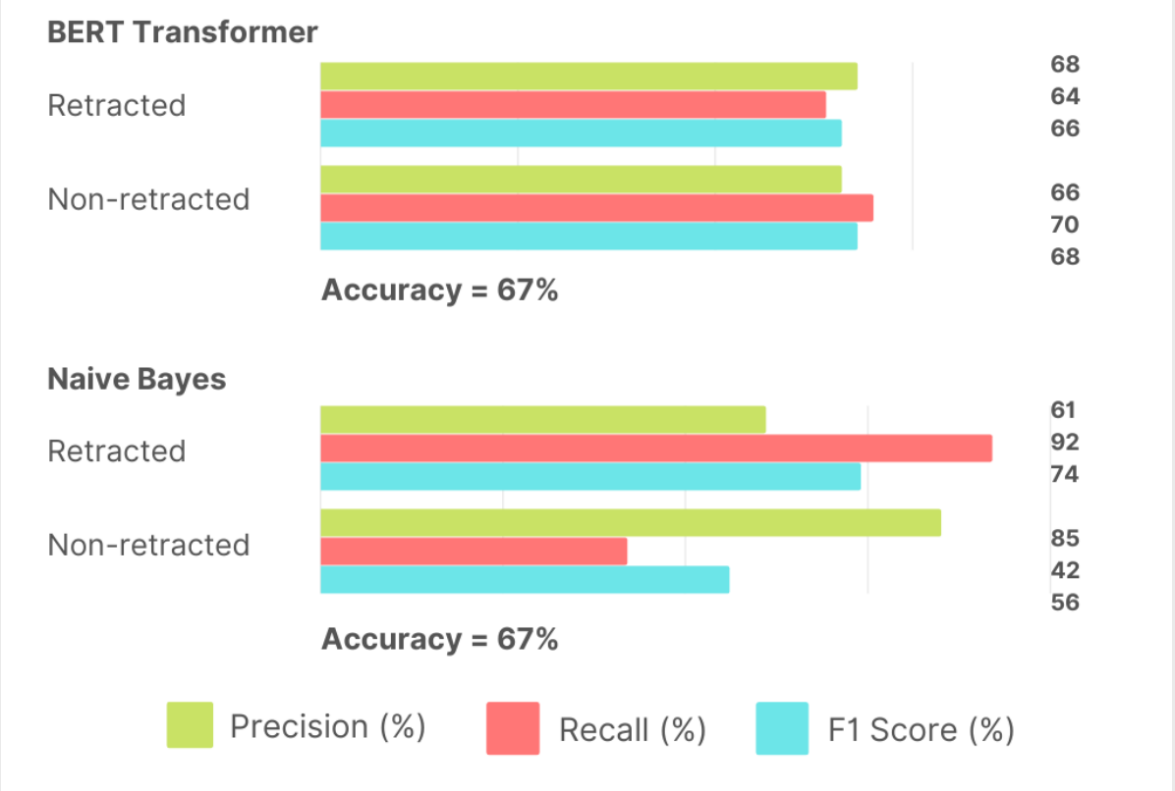
Detecting fraudulent research (Joep Franssen, Eveline Schmidt, Arleen Lindenmeyer, 2022)

Data: 1k retracted papers, 1k non-retracted papers from same journals

Method: Different methods (TF-IDF + NB, BERT)

Preprocessing: (1) none, (2) lemmatization, stop word removal, punctuation removal, lowercasing, and removal of white spaces (and tabs). We also removed proper nouns and deleted all numbers.

Results:



Applications in Economics and Finance

Article

Forecasting Net Income Estimate and Stock Price Using Text Mining from Economic Reports

Masahiro Suzuki [1,*], Hiroki Sakaji [1], Kiyoshi Izumi [1], Hiroyasu Matsushima [1] and Yasushi Ishikawa [2]

Suzuki, Sakaji, Izumi, Matsushima, Ishiwaka, 2020

Data: 17k reports on TOPIX securities from 5 brokers

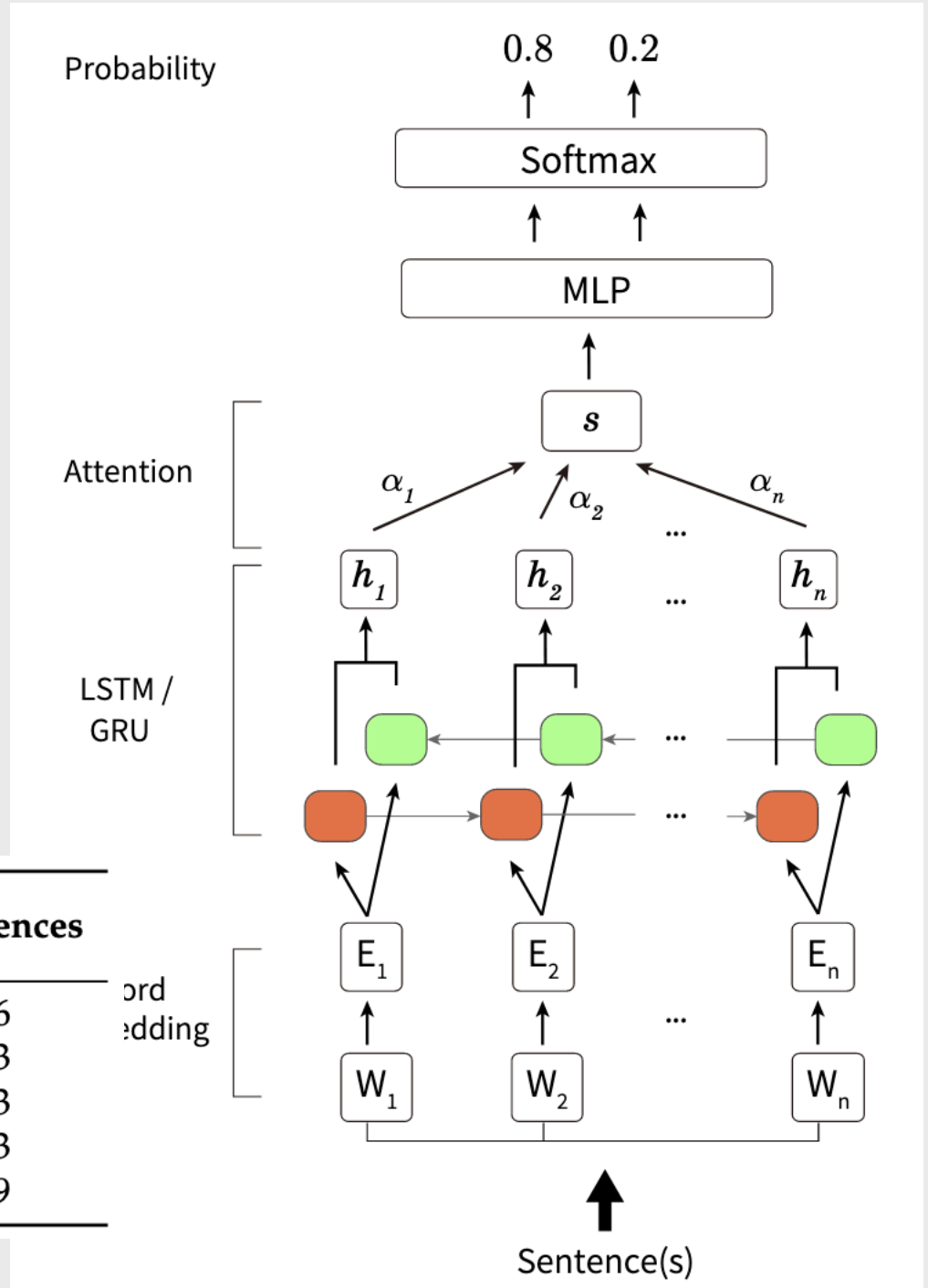
Method: LSTM/GRU + Self-attention

Preprocessing: decomposing sentences into words (the Japanese language does not have spaces between the words in a sentences)

Goal: Predict excess return (+/-) at 14 days

Results:

Broker	Opinion Sentences	Non-Opinion Sentences	Opinion and Non-Opinion Sentences	All Sentences
A	0.529	0.526	0.520	0.506
B	0.558	0.563	0.569	0.573
C	0.530	0.535	0.503	0.533
D	0.518	0.525	0.509	0.523
E	0.500	0.503	0.492	0.499



Applications in reports

Automatically code reports, detect emergencies
Extract safety risk factors

2018 IEEE International Conference on Software Quality, Reliability and Security

Identification of Security related Bug Reports via Text Mining using Supervised and Unsupervised Classification

Katerina Goseva-Popstojanova and Jacob Tyo
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV, USA
Email: Katerina.Goseva@mail.wvu.edu

Goseva-Popstojanova and Tyo, 2018

Data: 3,500 issue tracking systems of two NASA missions

Method: TF-IDF/TF + ML (Bayesian Network (BN), k-Nearest Neighbor (kNN), Naïve Bayes (NB), Naive Bayes Multinomial (NBM), Random Forest (RF), and Support Vector Machine (SVM))

Preprocessing: remove punctuation, characters and stop-words, stemming

Goal: Identify those software bugs reports that are security related

Results:

	Supervised System	*TFIDF_BN*	*TFIDF_kNN*	*TFIDF_NB*	*TFIDF_NBM*	*TFIDF_RF*	*TFIDF_SVM*
Flight Mission Developers Issues	*Accuracy*	68.0%	64.9%	**62.3%**	66.0%	70.6%	59.9%
	Precision	48.9%	66.7%	**73.4%**	65.9%	71.2%	73.0%
	Recall	94.4%	93.2%	**66.9%**	100.0%	92.9%	61.9%
	PFA	82.6%	89.7%	**46.7%**	99.5%	72.3%	44.0%
	F-Score	79.5%	77.7%	**70.0%**	79.5%	80.6%	67.0%
	G-Score	29.4%	18.5%	**59.3%**	1.0%	42.7%	58.8%

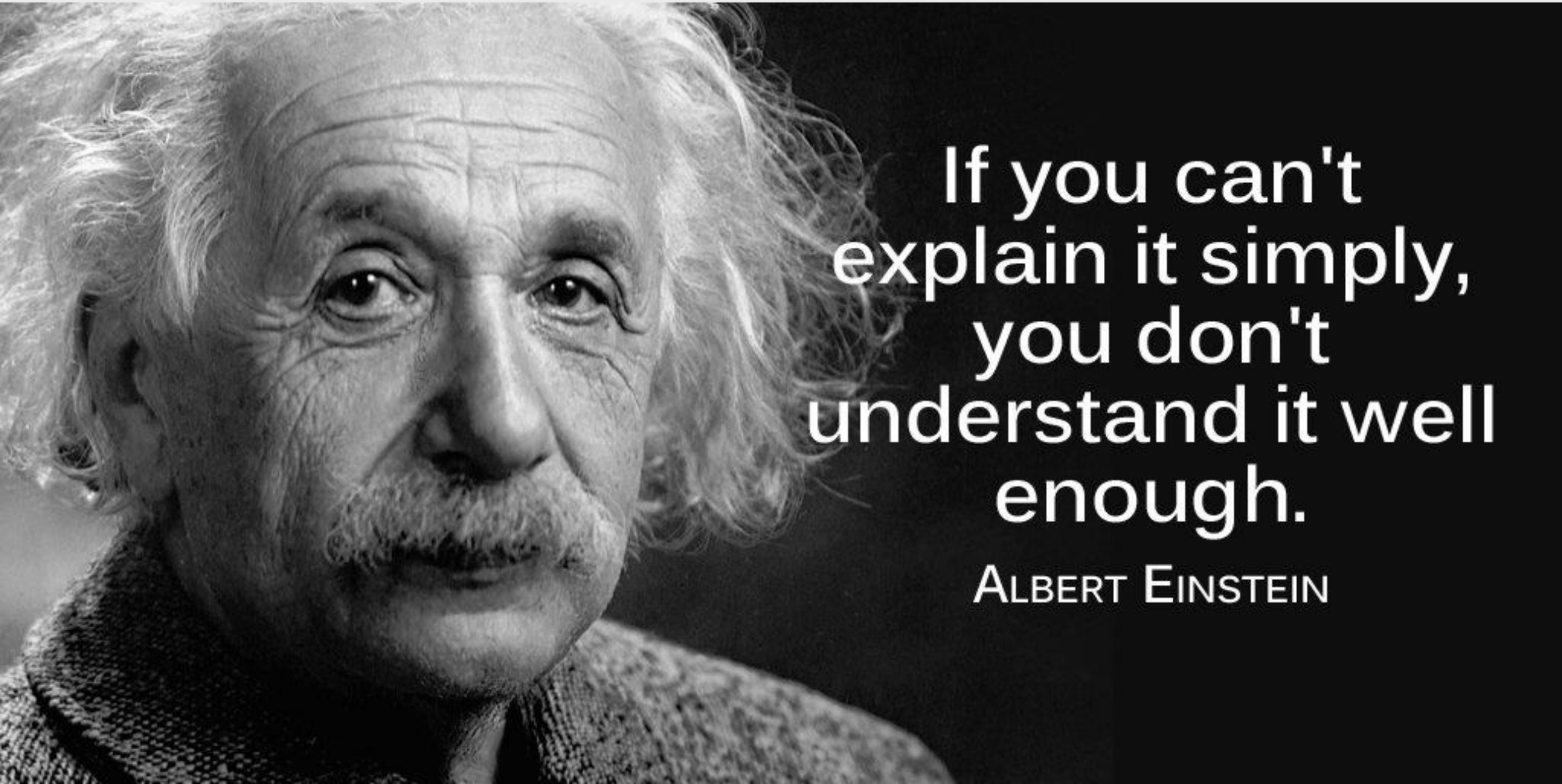
How to know if your results make sense?

Clustering: https://stats.stackexchange.com/questions/195456/how-to-select-a-clustering-method-how-to-validate-a-cluster-solution-to-warran

- Mixed methods (expertise is key)

Classification/Regression: Interpretability tools (next)

Interpretability



Interpretability

Being Right for the Right Reasons

Choices:

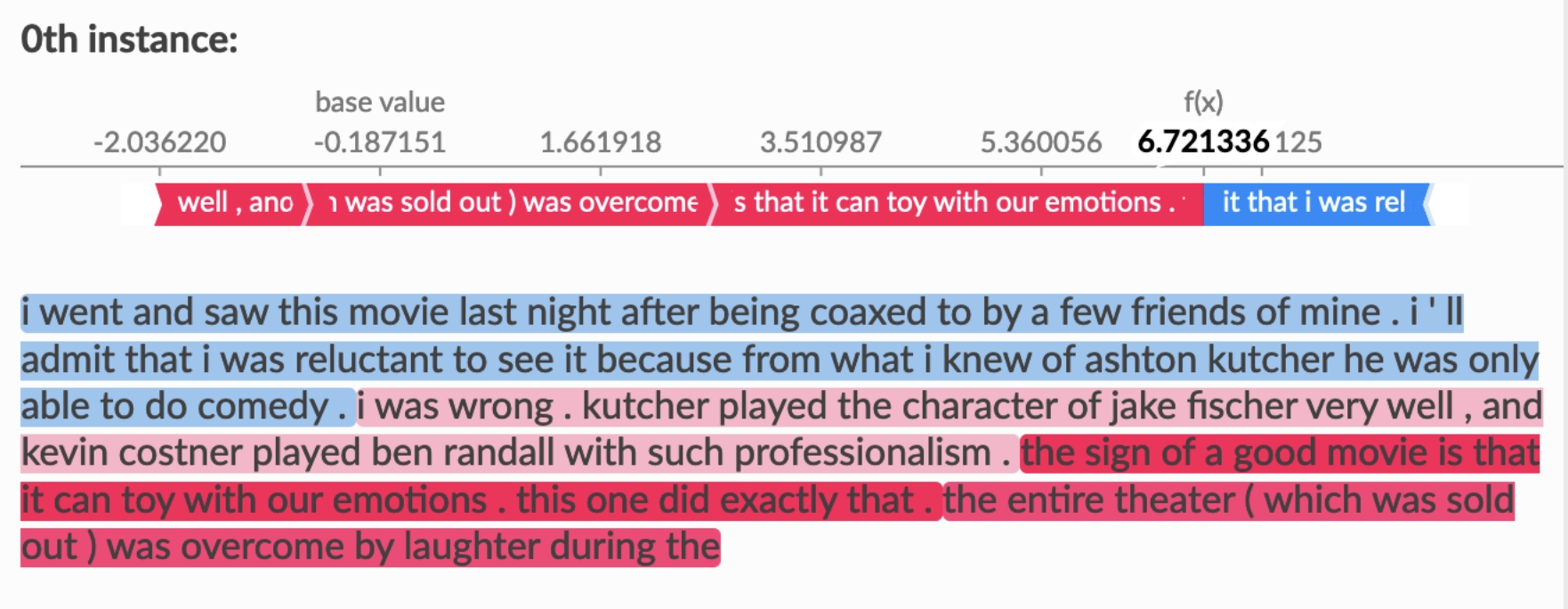
- Global vs Local interpretability
- Model-dependent vs model-agnostic interpretability

Local interpretability

Goal: Inform the human about the factors determining the prediction

Many advances in the last decade (SHAP, LIME, Anchors, Counterfactuals)

Basic idea: Change the observations slightly to observe how the prediction will change



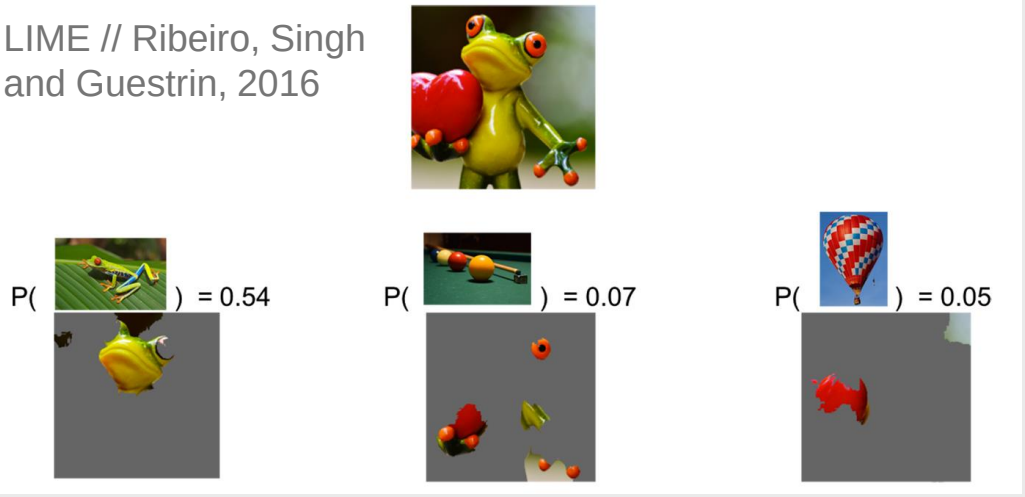
https://shap.readthedocs.io/en/latest/example_notebooks/text_examples/sentiment_analysis

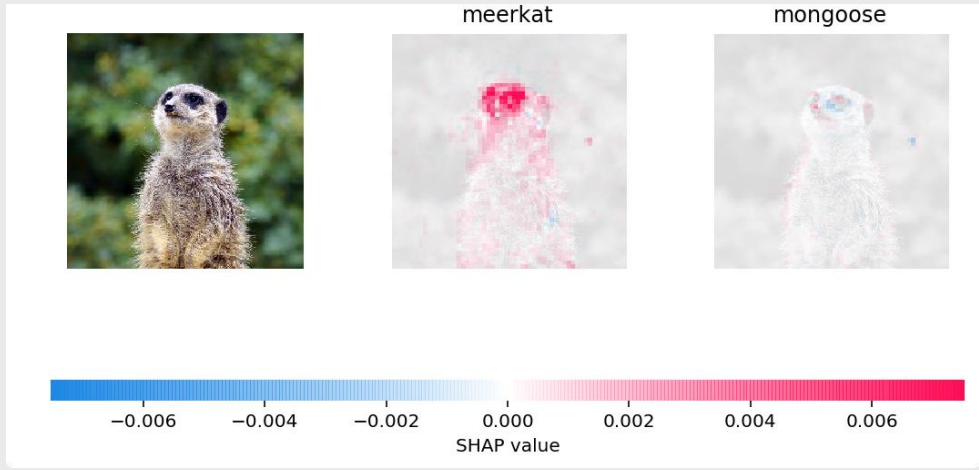
Local interpretability

Goal: Inform the human the reason of the prediction

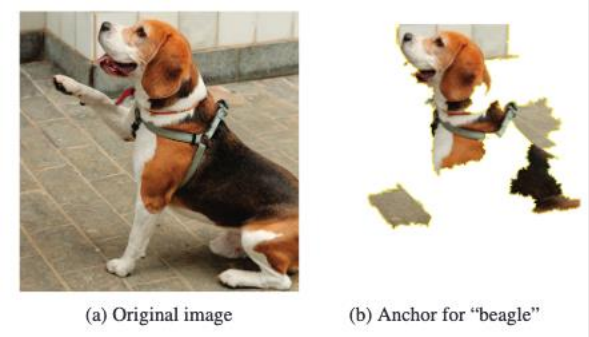
Basic idea: Change the observations slightly to observe how the prediction will change

LIME // Ribeiro, Singh and Guestrin, 2016

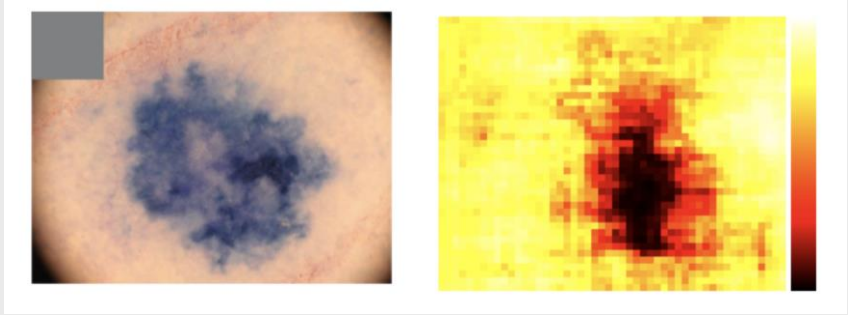




Scott M. Lundberg, Su-In Lee, 2017



Anchors // Ribeiro, Singh and Guestrin, 2018



https://silverpond.com.au/2018/04/17/an-ai-tells-us-what-it-knows-when-we-poke-it-in-the-eye/

Practical 8