Applications of text mining and NLP

Anastasia Giachanou

Summary

- Text -> Numeric
  - TF; TF-IDF; Word embeddings
- Similarity (often cosine similarity)
- Clustering
- LDA
- Classification
  - TF-IDF (+ Dimensionality reduction) + Classifier (e.g. LogisticRegression)
  - Word embeddings + Classifier (e.g. LogisticRegression)
  - Neural Networks (Feed-forward, RNN, LSTM, CNN, Transformers)

A collection of text mining applications

- Can you think of some text mining applications?

A collection of text mining applications

Similarity

Find authors of an anonymous book

Find duplicates and link records

Find relevant documents given a user query

Clustering

Targeted advertisement or learning

Recommendation systems (e.g. similar books)

Clustering stories (clustering fiction works, people's diagnoses, misinformation)

Track evolution of topics in discourse

Classification/Regression

Hate speech classification (similar: spam, fake news)

Sentiment and emotion analysis

Predict student performance

Probability of re-hospitalization

Classifying reports (e.g. hospital discharges, urgent issues)

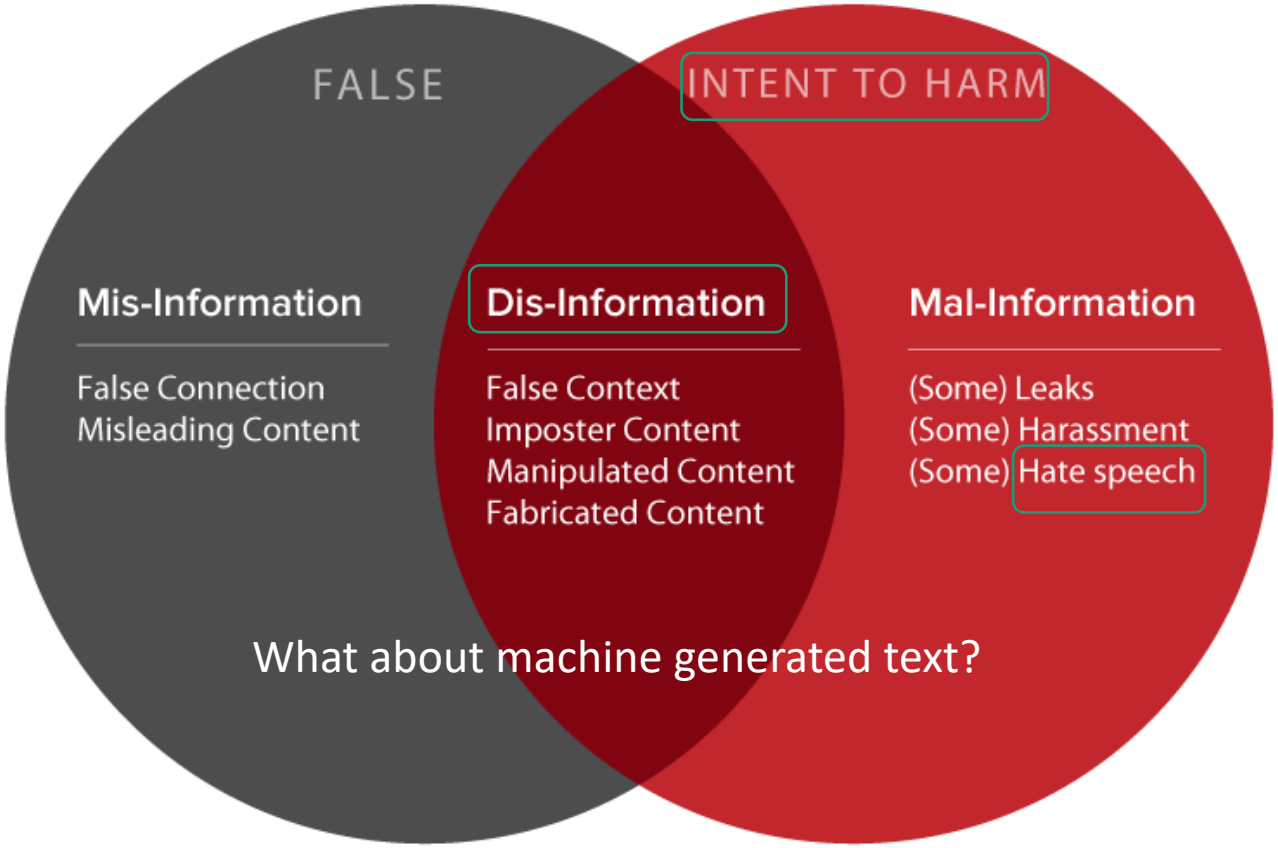
Predict stock market returns

Today

- Applications of text mining
  - Fake news detection
  - Hate speech detection
  - Text clustering in media
  - Healthcare applications
  - Interpretability

*50% of the presentation is based on a lecture given by Javier Garcia-Bernardo

Information disorder online



Council of Europe. https://www.coe.int/en/web/freedom-expression/information-disorder

Fake news detection

Information Credibility on Twitter

Carlos Castillo[1] Marcelo Mendoza[2,3] Barbara Poblete[2,4]

{chato,bpoblete}@yahoo-inc.com, marcelo.mendoza@usm.cl
[1]Yahoo! Research Barcelona, Spain
[2]Yahoo! Research Latin America, Chile
[3]Universidad Técnica Federico Santa María, Chile
[4]Department of Computer Science, University of Chile

Fake News Early Detection: A Theory-driven Model

XINYI ZHOU, ATISHAY JAIN, VIR V. PHOHA, and REZA ZAFARANI, Syracuse University, USA

DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning

Kashyap Popat[1], Subhabrata Mukherjee[2], Andrew Yates[1], Gerhard Weikum[1]
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Amazon Inc., Seattle, USA

Detection of conspiracy propagators using psycho-linguistic characteristics

Anastasia Giachanou
Universitat Politècnica de València, Spain; Utrecht University, The Netherlands

Bilal Ghanem
Universitat Politècnica de València, Spain; Symanto Research, Germany

Paolo Rosso
Universitat Politècnica de València, Spain

Definition of fake news

- A news article that is intentionally and verifiably false
  - where news broadly includes articles, claims, statements, speeches, posts, among other types of information related to public figures and organizations
- Fake news is intentionally false news published by a news outlet
  - emphasizes both news authenticity and intentions; it also ensures the posted information is news by investigating if its publisher is a news outlet

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19, 1 (2017), 22–36

Difficult to be detected by humans

- Human ability to detect deception is only slightly better than chance: accuracy rates are in the 55%-58% range
- Individuals trust fake news after repeated exposures (validity effect) or if it confirms their preexisting beliefs (confirmation bias) or if it pleases them (desirability bias)
- Peer pressure “controls” our perception and behavior (e.g., bandwagon effect)

Travel fast and more

- Research has shown that compared to the truth, fake news on Twitter is typically retweeted by many more users and spreads far more rapidly, especially for political news
- During the 2016 U.S. presidential election campaign, the top twenty frequently-discussed fake election stories generated **8,711,000** shares, reactions, and comments on Facebook, ironically, more than the **7,367,000** for the top twenty most-discussed election stories

S. Vosoughi, D. Roy, and S. Aral. (2018). The spread of true and false news online. Science 359, 6380, 1146–1151.
C. Silverman. (2016). This analysis shows how viral fake election news stories outperformed real news on Facebook. BuzzFeed News 16

The role of content

- Fake news potentially differ from the truth in terms of:
  - writing style and quality (by Undeutsch hypothesis)
  - quantity such as word counts (by information manipulation theory)
  - sentiments expressed (by four-factor theory)

U. Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. Handbuch der psychologie 11, 26–181
S. A McCornack, K. Morrison, J. E. Paik, A. M Wisner, and X. Zhu. 2014. Information manipulation theory 2: A propositional theory of deceptive discourse production. Journal of Language and Social Psychology 33, 4 (2014), 348–377
M. Zuckerman, B. M DePaulo, and R. Rosenthal. 1981. Verbal and Nonverbal Communication of Deception1. In Advances in experimental social psychology. Vol. 14. Elsevier, 1–59

Information credibility on Twitter

- Assessing the credibility of a given set of tweets
- Data collection: collected all the tweets matching queries during a 2-days window; 2,500 topics; Amazon Mechanical Turk
- Features:
  - Message-based: tweet length, existence of special chars, sentiment, etc.
  - User-based: registration age, number of followers, followees, etc.
  - Topic-based: the fraction of tweets with URL, fraction of positive, etc.
  - Propagation-based: depth of a re-tweet, number of initial tweets of a topic
- Decision Tree classifier

C. Castillo, M. Mendoza, and B. Poblete. (2011). Information credibility on Twitter. In Proceedings of the 20th international conference on World wide web, pp. 675-684

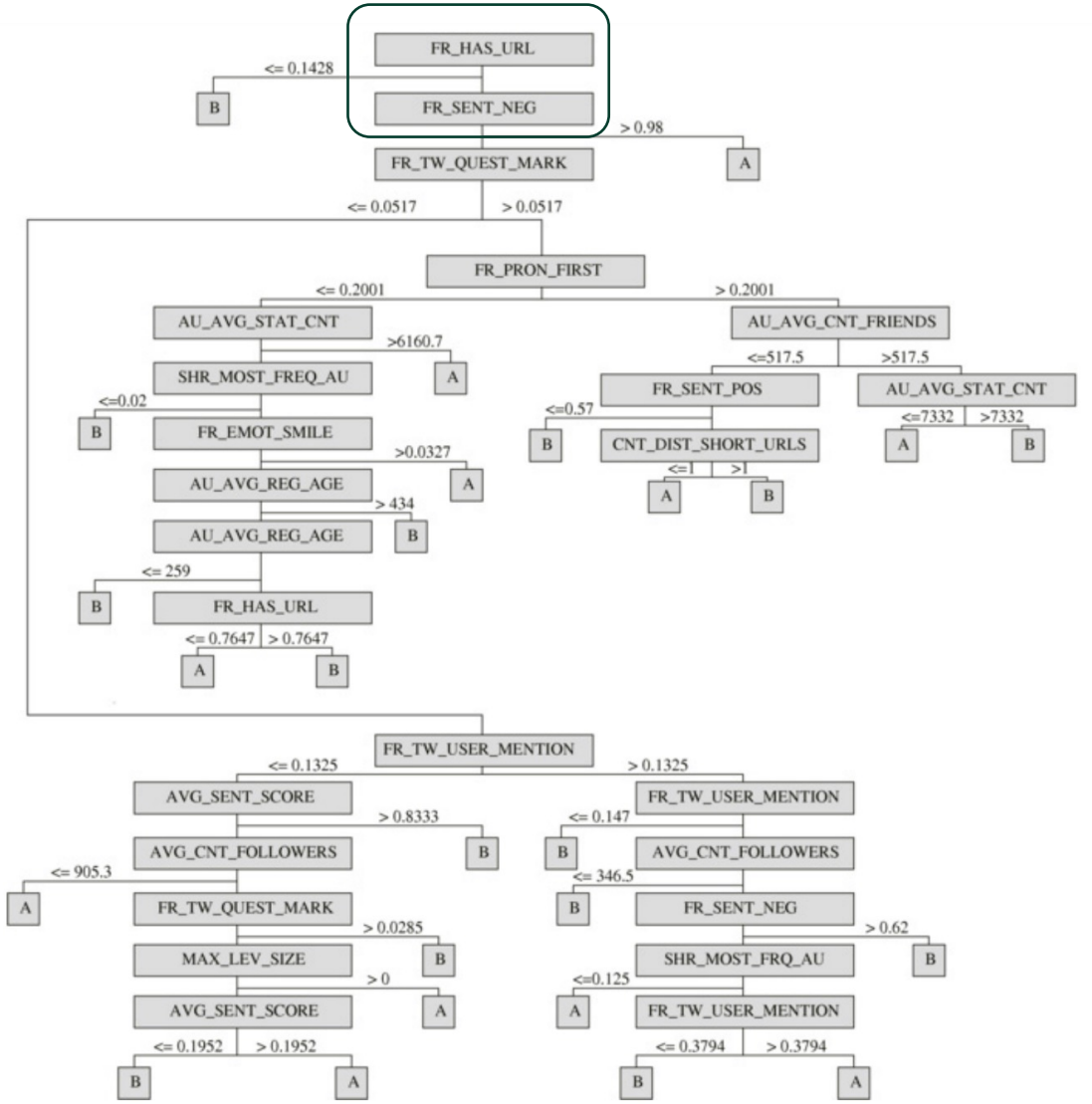


Table 8: Experimental results obtained for the classification of credibility cases. The training step was conducted using four different subsets of features.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
Text subset					
A	0.636	0.152	0.808	0.636	0.712
B	0.848	0.364	0.700	0.848	0.767
W. Avg.	0.742	0.258	0.754	0.742	0.739
Network subset					
A	0.667	0.212	0.759	0.667	0.71
B	0.788	0.333	0.703	0.788	0.743
W. Avg.	0.727	0.273	0.731	0.727	0.726
Propagation subset					
A	0.606	**0.091**	**0.870**	0.606	0.714
B	**0.909**	0.394	0.698	**0.909**	0.789
W. Avg.	0.758	0.242	0.784	0.758	0.752
Top-element subset					
A	**0.727**	0.152	0.828	**0.727**	**0.774**
B	0.848	**0.273**	**0.757**	0.848	**0.800**
W. Avg.	0.788	0.212	0.792	0.788	**0.787**

C. Castillo, M. Mendoza, and B. Poblete. (2011). Information credibility on Twitter. In Proceedings of the 20th international conference on World wide web, pp. 675-684

Fake news early detection: A theory-driven model

- Features:
  - Lexicon-level (e.g., BoW)
  - Syntax-level (e.g., POS)
  - Semantic-level (e.g., General Clickbait Patterns, Readability, Sensationalism, News-worthiness)
  - Discourse-level (Rhetorical Relationships)
- Several supervised classifiers with five-fold cross-validation
- SVM, Random Forest, and XGBoost perform best
- PolitiFact and BuzzFeed

X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani. (2019). Fake News Early Detection: A Theory-driven Model. arXiv preprint arXiv:1904.11679

Content Quality

	Feature(s)	Example	Tool & Ref.
Informality	#/% Swear Words	"damn"	Linguistic Inquiry and Word Count (LIWC)
	#/% Netspeak	"btw"	
	#/% Assent	"OK"	
	#/% Nonfluencies	"umm"	
	#/% Fillers	"you know"	
	Overall #/% Informal Words	/	
Subjectivity	#/% Biased Lexicons	"attack"	[1]
	#/% Report Verbs	"announce"	
	#/% Factive Verbs	"observe"	[2]
Diversity	#/% Unique Words	/	/
	#/% Unique Content Words	"car"	LIWC
	#/% Unique Nouns	/	POS Taggers
	#/% Unique Verbs	/	
	#/% Unique Adjectives	/	
	#/% Unique Adverbs	/	

Quantity

Characters
Words
Sentences
Paragraphs
Avg. # Characters Per Word
Avg. # Words Per Sentence
Avg. # Sentences Per Paragraph

Cognitive Process

#/% Insight	"think"	LIWC
#/% Causation	"because"	
#/% Discrepancy	"should"	
#/% Tentative	"perhaps"	
#/% Certainty	"always"	
#/% Differentiation	"but"	
Overall #/% Cognitive Processes		

Sentiment

#/% Positive Words	LIWC
#/% Negative Words	
#/% Anxiety Words	
#/% Anger Words	
#/% Sadness Words	
Overall #/% Emotional Words	
Avg. Sentiment Score of Words	NLTK

Perceptual Process

#/% See	LIWC
#/% Hear	
#/% Feel	
Overall #/% Perceptual Processes	

X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani. (2019). Fake News Early Detection: A Theory-driven Model. arXiv preprint arXiv:1904.11679

Disinformation-related Attribute(s)	PolitiFact				BuzzFeed			
	XGBoost		RF		XGBoost		RF	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Quality	.667	.652	.645	.645	.556	.500	.512	.512
– Informality	.688	.727	.604	.604	.555	.513	.508	.508
– Subjectivity	.688	.706	.654	.654	.611	.588	.533	.530
– Diversity	.583	.600	.620	.620	.639	.552	.544	.544
Sentiment	.625	.591	.583	.583	.556	.579	.515	.525
Quantity	.583	.524	.638	.638	.528	.514	.584	.586
Specificity	.625	.609	.558	.558	.583	.571	.611	.611
– Cognitive Process	.604	.612	.565	.565	.556	.579	.531	.531
– Perceptual Process	.563	.571	.612	.612	.556	.600	.571	.571
Overall	**.729**	**.735**	**.755**	**.755**	**.667**	**.647**	**.625**	**.625**

Individual attributes perform similarly, while combining all attributes performs better in predicting fake news.

Rank	PolitiFact		BuzzFeed	
	Feature	Attribute	Feature	Attribute
1	# Characters per Word	Quantity	# Overall Informal Words	Informality
2	# Sentences per Paragraph	Quantity	% Unique Words	Diversity
3	% Positive Words	Sentiment	% Unique Nouns	Diversity
4	% Unique Words	Diversity	% Unique Content Words	Diversity
5	% Causation	Cognitive Process	# Report Verbs	Subjectivity
6	# Words per Sentence	Quantity	% Insight	Cognitive Process
7	% Report Verbs	Subjectivity	% Netspeak	Informality
8	% Unique Verbs	Diversity	# Sentences	Quantity
9	# Sentences	Quantity	% Unique Verbs	Diversity
10	% Certainty Words	Cognitive Process	% Unique Adverbs	Diversity

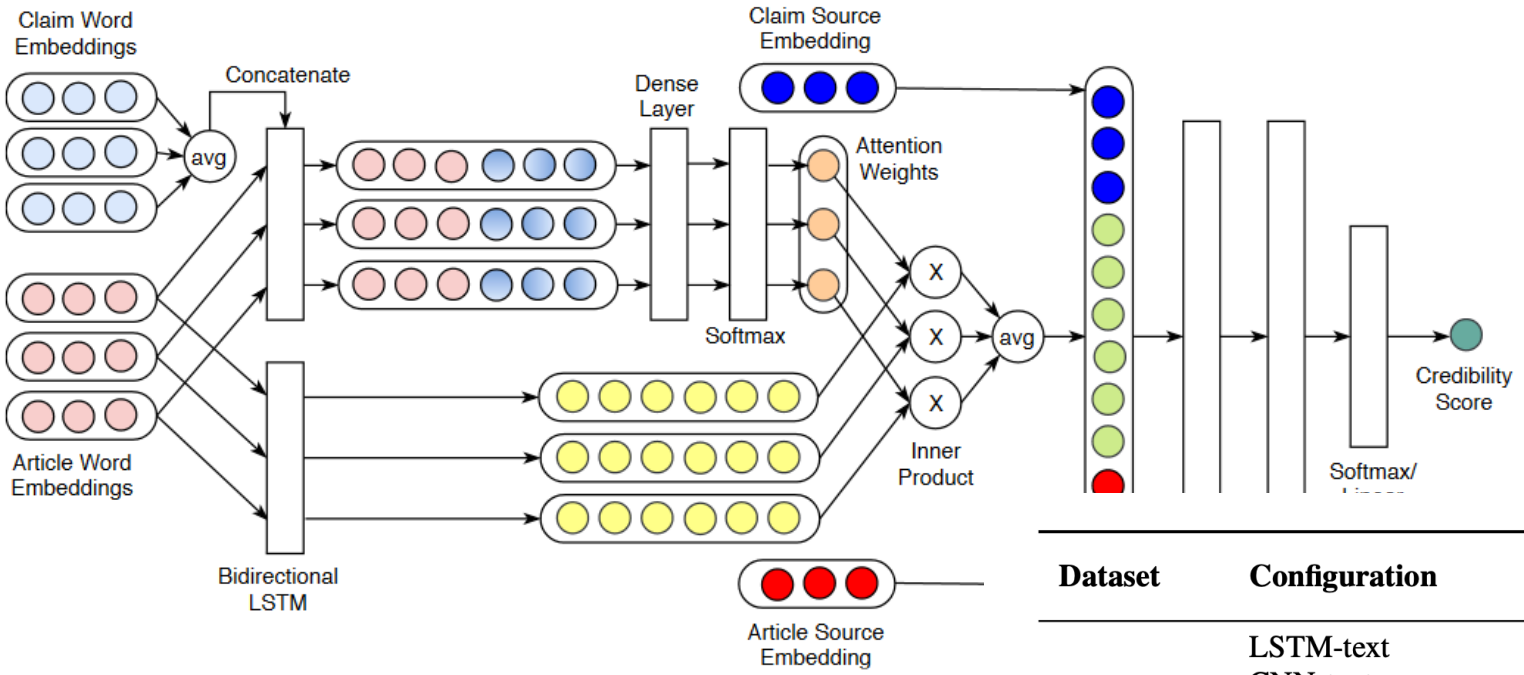
In both datasets, content diversity and quantity are most significant in differentiating fake news from the truth; cognitive process involved and content subjectivity are second; content informality and sentiments expressed are third.

X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani. (2019). Fake News Early Detection: A Theory-driven Model. arXiv preprint arXiv:1904.11679

Declare: Debunking fake news and false claims using evidence-aware deep learning

- Credibility of arbitrary claims made in natural language text
- Data collection:
  - Snopes: 4341 claims;PolitiFact: 3568 claims; NewsTrust: 5344 claims; RumorEval-2017: 272 claims
- Each claim as a query to BING search engine and retrieve the top 30 search results with their respective web sources

K. Popat, S. Mukherjee, A. Yates, and G. Weikum. (2018). DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 22-32



Dataset	Configuration	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro F1-Score	AUC
Snopes	LSTM-text	64.65	64.21	0.66	0.70
	CNN-text	67.15	63.14	0.66	0.72
	Distant Supervision	**83.21**	**80.78**	**0.82**	**0.88**
	DeClarE (Plain)	74.37	78.57	0.78	0.83
	DeClarE (Plain+Attn)	78.34	78.91	0.79	0.85
	DeClarE (Plain+SrEmb)	77.43	79.80	0.79	0.85
	DeClarE (Full)	78.96	78.32	0.79	0.86
PolitiFact	LSTM-text	63.19	61.96	0.63	0.66
	CNN-text	63.67	63.31	0.64	0.67
	Distant Supervision	62.53	62.08	0.62	0.68
	DeClarE (Plain)	62.67	69.05	0.66	0.70
	DeClarE (Plain+Attn)	65.53	68.49	0.66	0.72
	DeClarE (Plain+SrEmb)	66.71	69.28	0.67	0.74
	DeClarE (Full)	**67.32**	**69.62**	**0.68**	**0.75**

K. Popat, S. Mukherjee, A. Yates, and G. Weikum. (2018). DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 22-32

Read this before they delete it: A psycho-linguistic analysis of conspiracy theorists

Table 1: Hashtags used to collect the tweets and statistics about the collection.

	pro-conspiracy	anti-conspiracy
Hashtags	#vaccinesCauseAutism #antiVax #climateChangeIsNotReal #flatEarth #nasaLies #nasaFake #spaceIsFake #moonLandingFake #bigPharmaFraud #ebolaconspiracy #antiFluoridation	#vaccinesWork #vaccinessavelives #climateChangeIsReal #earthisnotflat #nasatruth #nasaIRreal #spaceIsReal #moonlandingisreal
users	977	950
tweets	912,735	992,798

Opponents have:

- old and verified accounts
- a larger number of statuses
- higher usage of work, leisure, money, home, and death, causation (because, effect, hence)

Supporters have:

- less followers, less statuses, favorites and friends
- concern more about religion
- use more swear words

Giachanou, A, Ghanem, B. , Rosso, P. "Detection of conspiracy propagators using psycho-linguistic characteristics." *Journal of Information Science* 49.1 (2023): 3-17.

Read this before they delete it: A psycho-linguistic analysis of conspiracy theorists

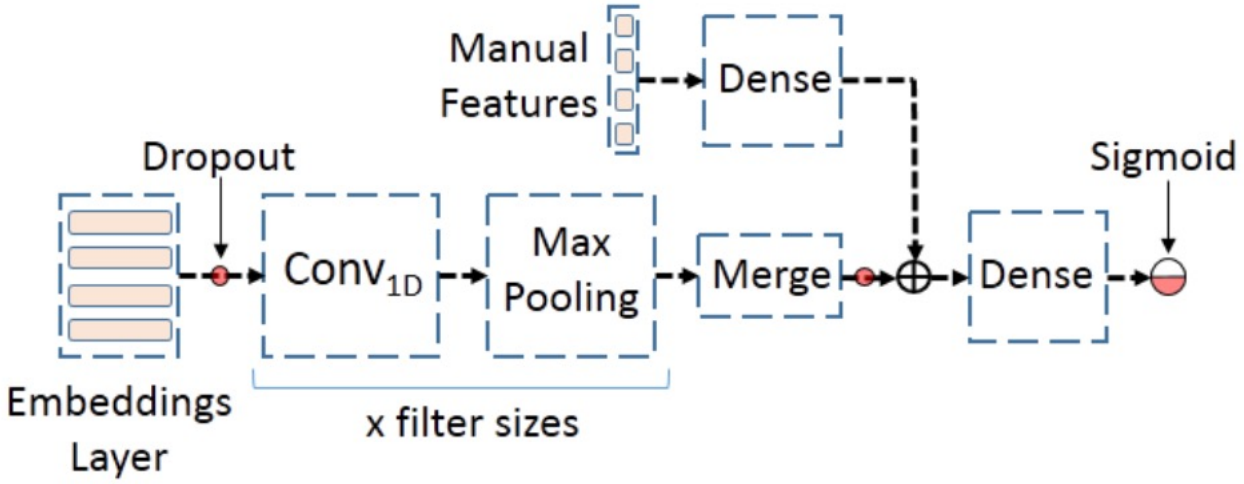


Fig. 1. Architecture of ConspiDetector.

Majority class	0.34
Random	0.50
USE	0.69
CNN	0.68
CNN + Profile	0.58
CNN + Personality	0.73
CNN + LIWC	0.71
CNN + Sentiment	0.66
CNN + Emotion	0.67
ConspiDetector (psycho-linguistic)	**0.74**
CNN + Psycho-linguistic + Profile	0.68

Giachanou, A, Ghanem, B. , Rosso, P. "Detection of conspiracy propagators using psycho-linguistic characteristics." *Journal of Information Science* 49.1 (2023): 3-17.

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

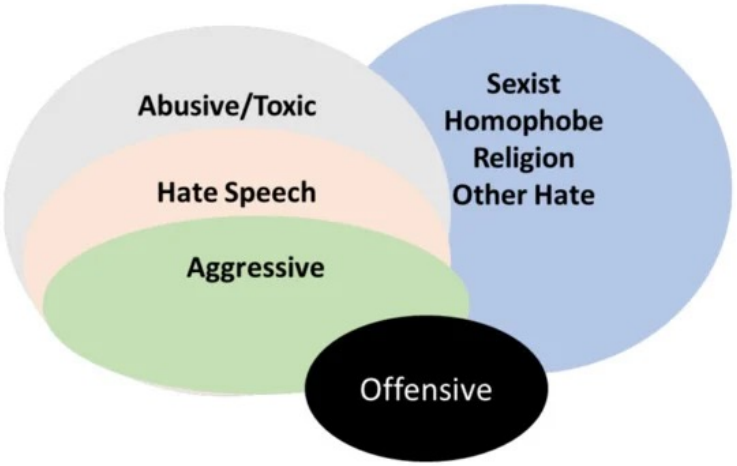
Zeerak Waseem, James Thorne and Joachim Bingel

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

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

Hate Speech

- Social media enable the propagation of hate speech
- Hate speech detection is crucial to reducing crime and protecting people's beliefs
  - On July 13, D66 leader Sigrid Kaag announced her decision not to continue in politics via Twitter. She mentions "hate, intimidation and threats" in the statement and the effect on her family is the reason to stop.
  - https://nos.nl/nieuwsuur/artikel/2482833-toelichting-twitter-reacties-op-vertrek-sigrid-kaag



Hate Speech Pyramid

- The 2019 UN Strategy and Plan of Action on Hate Speech
- '**Attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor**'.
- Genocidal acts cannot occur without being upheld by the lower stages that act as a base for mass atrocities.

Hateful Symbols

Data: Annotation of 16k tweets based on Gender studies and Critical Race Theory (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.

Z. Waseem and D. Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93,. ACL.

Using CNN for Hate Sppech

Data: Waseem&Hovy (2016)

Method: CNN using word embeddings and character n-grams

Word embeddings: word2vec and random vectors

Preprocessing: None

Classifier: Softmax

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)

B. Gambäck and U. Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, ACL.

Multi Task Learning for Hate Speech

Aim: Train a model that is robust across data originating from different distributions and labeled under differing annotation guidelines

Data:

-Waseem & Hovy (2016), 25k annotated tweets, 11-point test based on work in the fields of Gender Studies and CRT, no geographical restriction, (racism, sexism, neither, and both)

-Davidson et al, 2017; Target groups; Twitter guidelines; US data;(hate speech, offensive, and neither)

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

Feed-forward neural network with 2 hidden layers

Preprocessing: Removing usernames, links and punctuation

Classifier: Softmax

Results:

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online harassment*, 29-55.

Training objective		Features	F_1 -scores of predictions on test sets							
Primary task	Auxiliary task		W/W+H				Davidson			
			Racism	Sexism	Neither	Average	Hate speech	Offensive	Neither	Average
W/W+H	–	BoW	0.70	0.65	0.88	0.82	0.00	0.64	0.42	0.57
W/W+H	–	Emb	0.30	0.42	0.85	0.71	0.01	0.04	0.29	0.08
W/W+H	–	B+E	0.00	0.00	0.82	0.57	0.00	0.00	0.29	0.05
Davidson	–	BoW	0.22	0.29	0.69	0.56	0.32	0.94	0.84	0.89
Davidson	–	Emb	0.00	0.32	0.60	0.48	0.19	0.92	0.69	0.84
Davidson	–	B+E	0.25	0.33	0.70	0.58	0.39	0.82	0.94	0.89
Both	–	BoW	0.21	0.54	0.81	0.70	0.20	0.92	0.77	0.86
Both	–	Emb	0.21	0.45	0.76	0.64	0.05	0.90	0.64	0.80
~~Both~~	~~–~~	~~B+E~~	~~0.17~~	~~0.53~~	~~0.81~~	~~0.69~~	~~0.31~~	~~0.92~~	~~0.77~~	~~0.86~~
W/W+H	Davidson	BoW	0.64	0.63	0.87	0.80	0.39	0.94	0.84	0.89
W/W+H	Davidson	Emb	0.32	0.50	0.84	0.72	0.10	0.91	0.64	0.82
W/W+H	Davidson	B+E	0.51	0.53	0.86	0.75	0.16	0.93	0.78	0.86
Davidson	W/W+H	BoW	0.66	0.62	0.86	0.79	0.37	0.94	0.83	0.89
Davidson	W/W+H	Emb	0.39	0.49	0.84	0.73	0.09	0.91	0.62	0.81
Davidson	W/W+H	B+E	0.60	0.57	0.85	0.77	0.14	0.93	0.78	0.86

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. *Online harassment*, 29-55.

BERT for Hate Speech

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

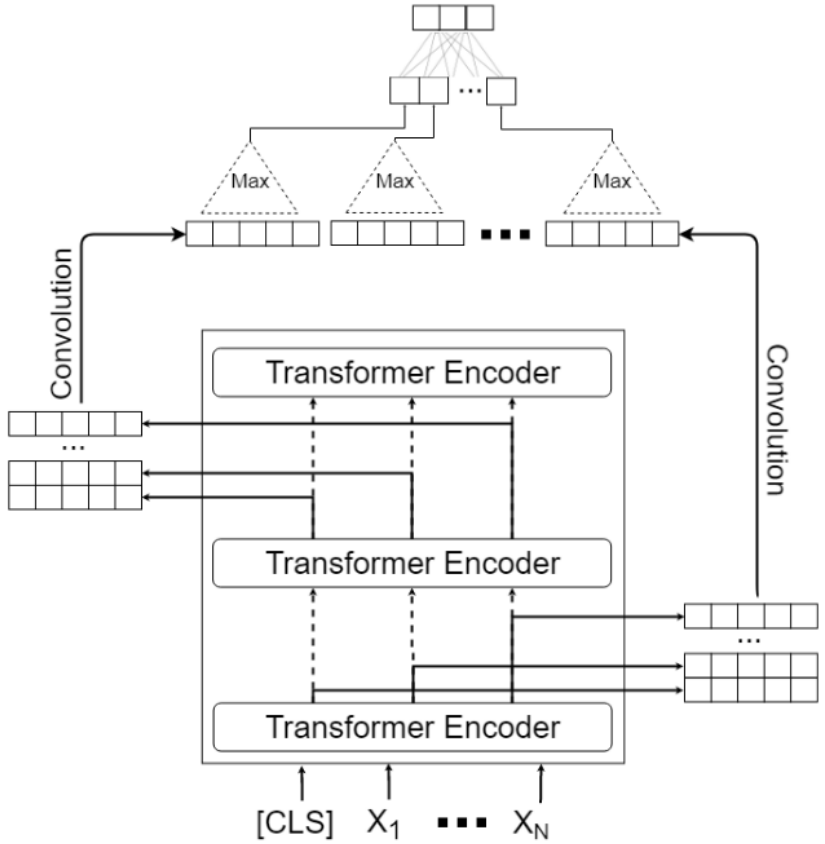
Best method: BERT + CNN

Each layer of the transformer gives an output → CNN

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

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**



Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. In *the 8th International Conference on Complex Networks and Their Applications.* (pp. 928-940). Springer.

BERT for Hate Speech

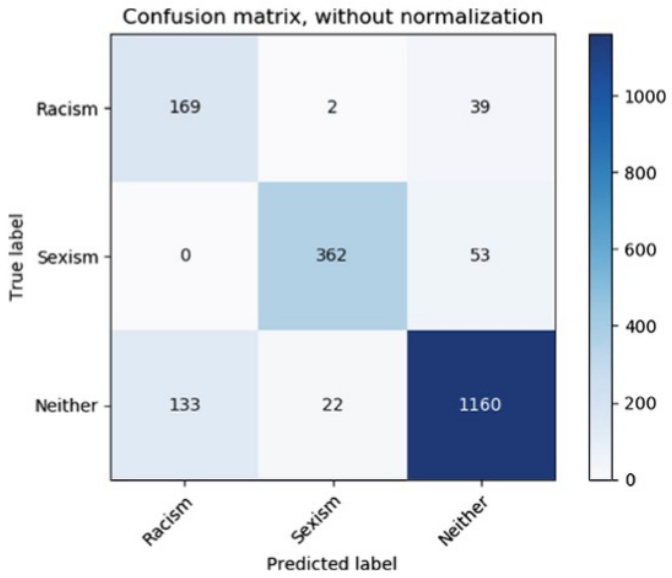


Fig. 2. Waseem-datase's confusion matrix

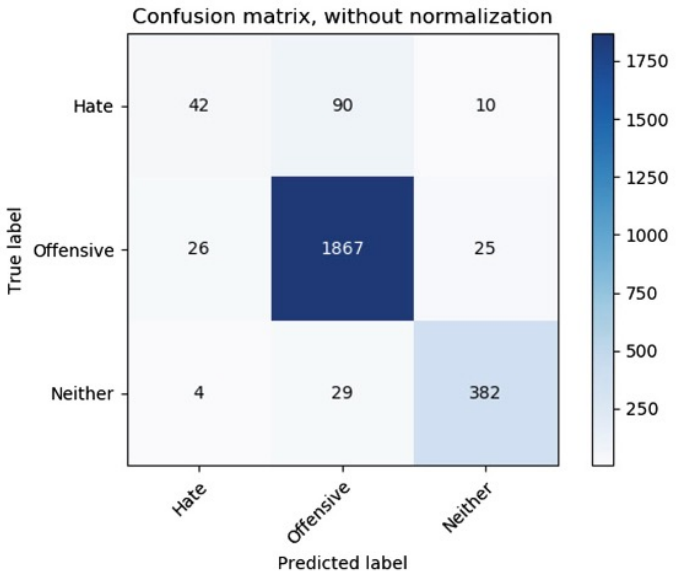


Fig. 3. Davidson-dataset's confusion matrix

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. In *the 8th International Conference on Complex Networks and Their Applications.* (pp. 928-940). Springer.

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:

- To what extent is the WAR figurative frame and the conventional metaphor DISEASE TREATMENT IS WAR used to talk about Covid-19 on Twitter?

- Which lexical units are used within this metaphorical frame and which lexical units are not?

- 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) - 25.000 tweets per day

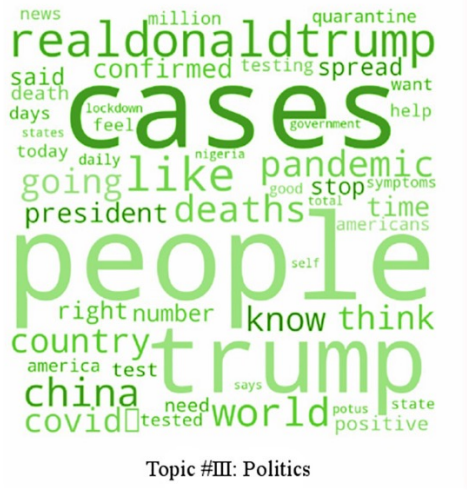
Method: LDA (4 and 16 topics) + correlation of topics with frames

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

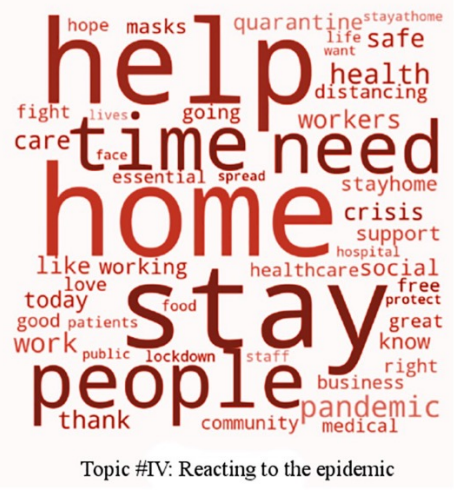
COVID pandemic (Wicke & Bolognesi 2020)

Topic #I: Communications and Reporting

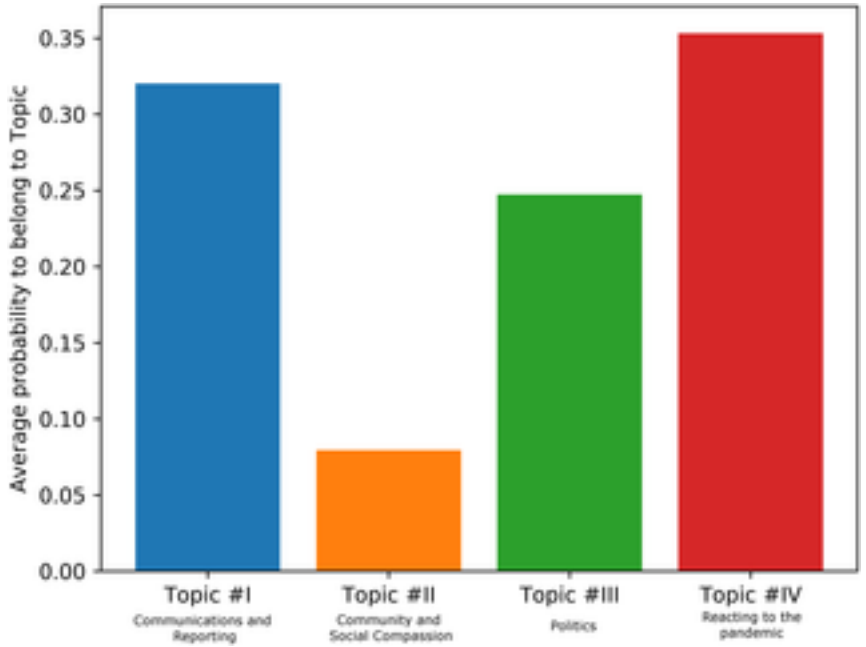
Topic #II: Community and Social Compassion



Topic #III: Politics



Topic #IV: Reacting to the epidemic



LDA-predicted average probability of WAR term contributing to one of 4 topics.

The results show that 5.32% of all tweets contain war-related terms

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

Data: 130k articles from 24 news outlets

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

Preprocessing: Unclear

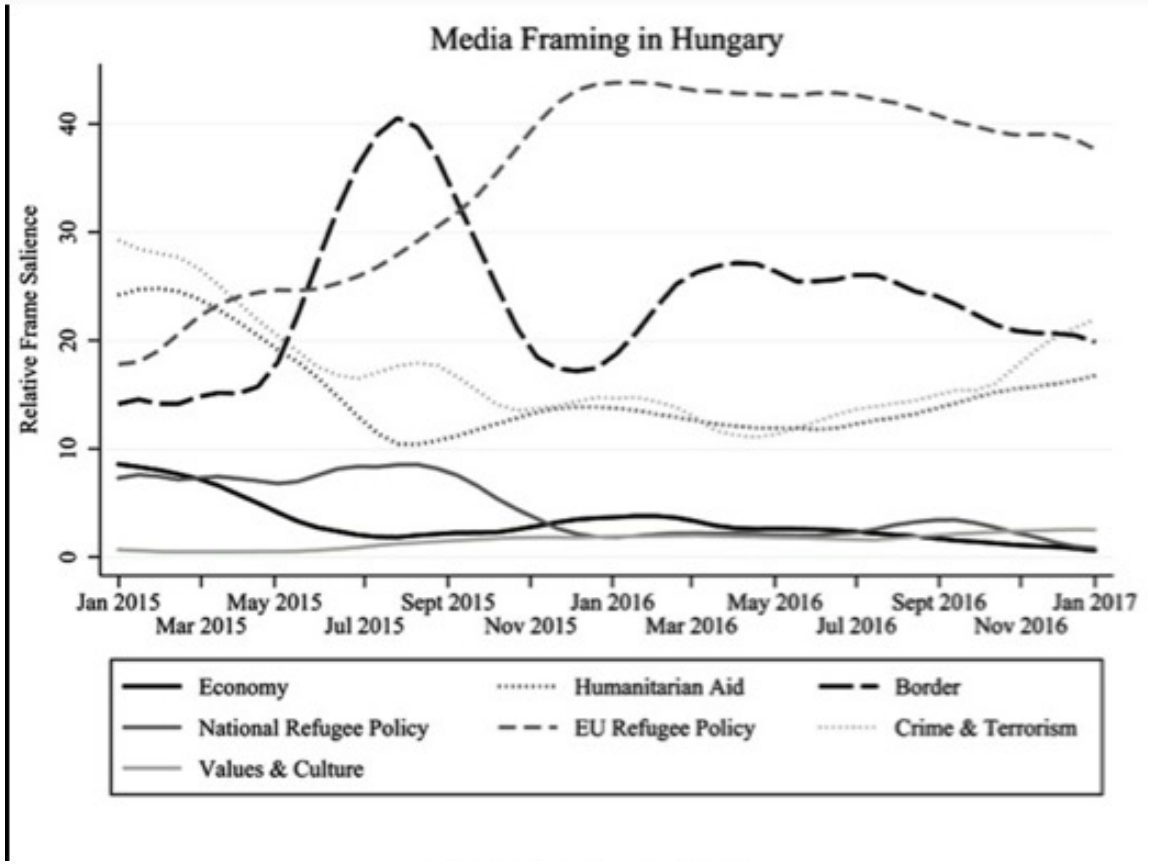
Validation: Semantic validity (are the topics distinctive) + Randomly reading three articles per topic/country + predictive validity (are important events such as elections reflected)

Media Corpora Description

Country	Media outlets	Keywords	*N* (articles)
Hungary	*Magyar Hirlap*, *Magyar Idők*, *Nepszabadsag*, *Nepszava*	menedék* or menekült*	8,865
Germany	*BILD*, *Frankfurter Rundschau*, *Spiegel Online*, *taz*, *Welt Online*, *ZEIT Online*	asyl* or flüchtling*	58,526
Sweden	*Aftonbladet*, *Dagens Industri*, *Dagens Nyheter*, *Expressen*, *Svenska Dagbladet*	asyl* or flykting*	17,789
United Kingdom	*Daily Mirror*, *The Daily Telegraph*, *The Guardian*, *Metro, mirror.co.uk, telegraph.co.uk*	asyl* or refugee*	31,223
Spain	*ABC*, *El Mundo*, *El Pais*	asilo* or refugiad*	13,639

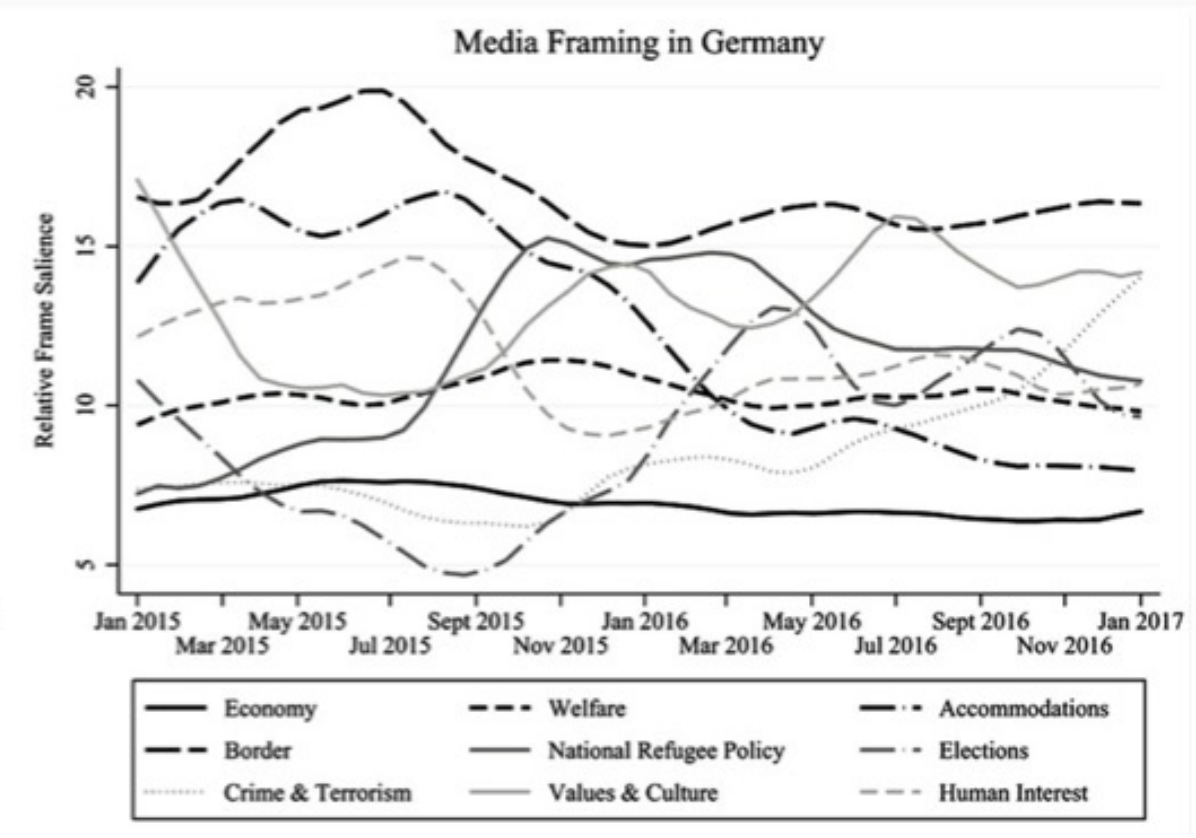
Refugees crisis (Heidenreich, Lind, Eberl & Boomgaarden, 2019)

- Before May 2015, media framing was mainly concerned with (international) ‘humanitarian aid’
- In October 2015, the Hungarian government decided to close its border to Croatia and the framing shifted to the European level (i.e. ‘EU refugee policy’)



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

- After Merkel's well-known assertion '*Wir schaffen das*' on 31 August, the 'national refugee policy', the question of how to deal with refugees now that they are in the country becomes more relevant
- The search for 'accommodations' plays a particularly important role in German media
- The 'crisis' also played an important role in the regional elections in March and September 2016



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**

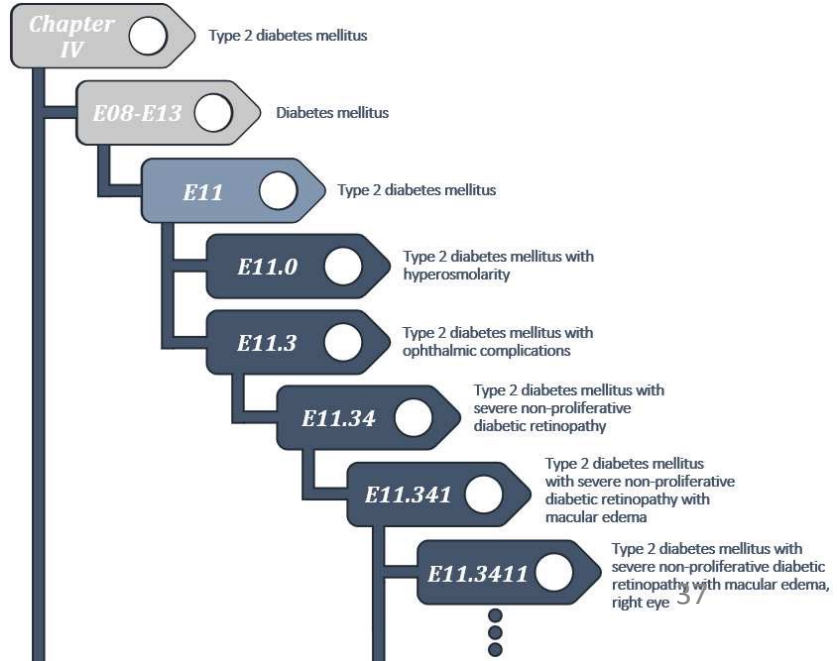
Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks

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

ICD-10 coding

- Medical coding is used to identify and standardize clinical concepts in the records collected from healthcare services
- The ICD- 10 is the most widely-used coding with more than 11,000 different diagnoses, affecting research, reporting, and funding



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**

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:

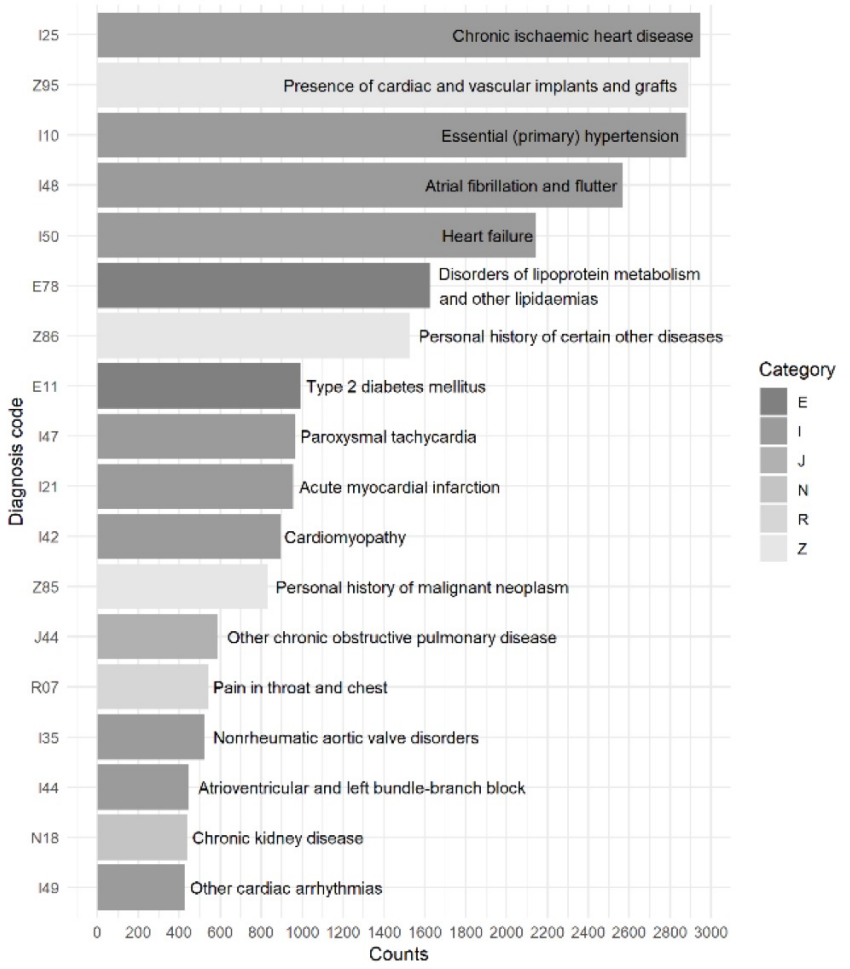


Figure 1: ICD rolled-up codes with more than 400 appearances in the UMCU dataset.

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

Table 2: Single-label performance: accuracy and *F1* score on two settings (ICD chapters and rolled-up ICDs) for the models when trained on the UMCU discharge letters.

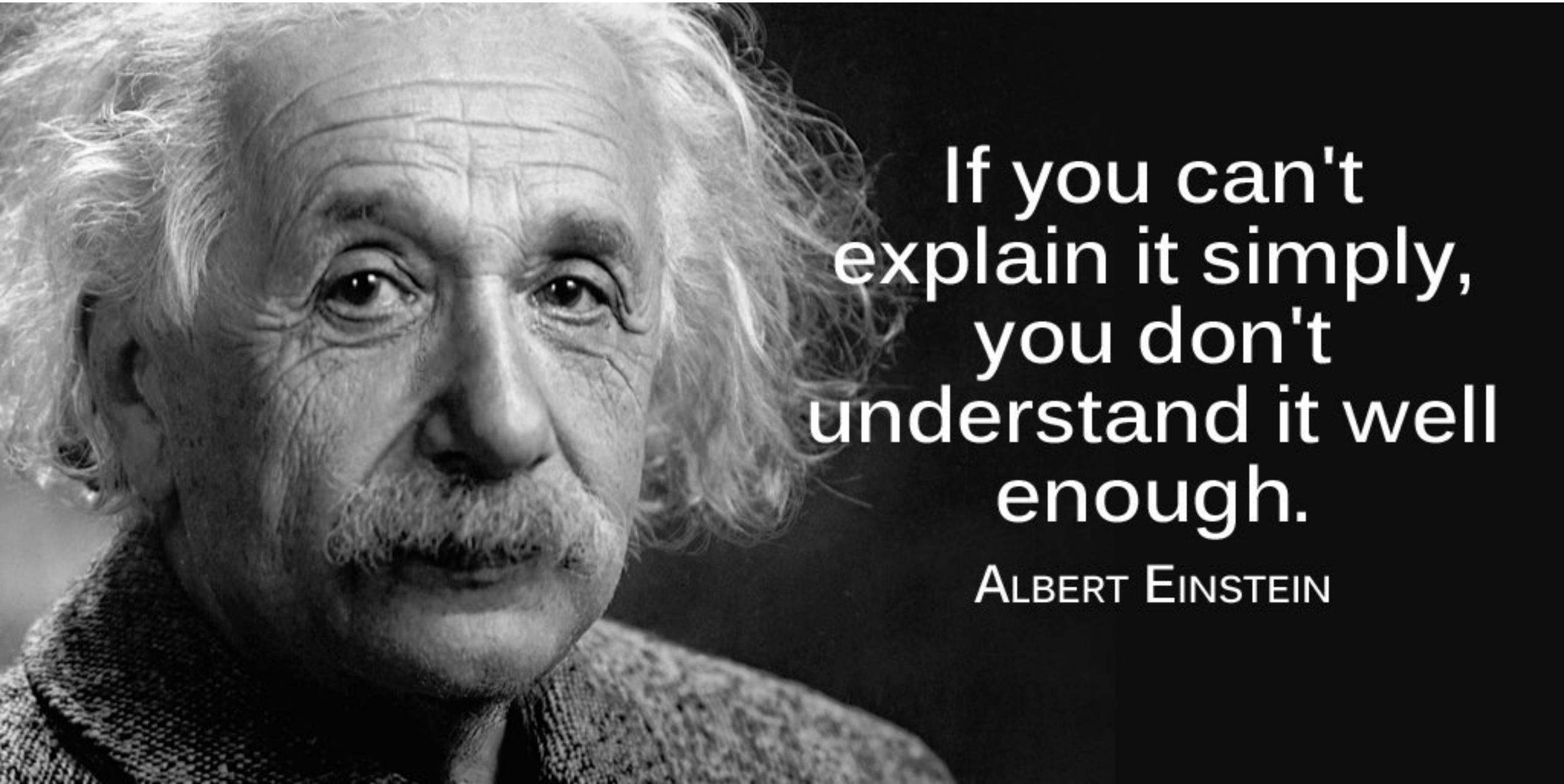
	ICD chapters		**Rolled-up ICD codes**	
	Accuracy	F1	Accuracy	F1
BOW SVM (baseline)	54.8	54.8	14.1	14.1
Average word embeddings (SVM)	54.9	**54.9**	18.2	18.2
CNN(1conv)	57.3	49.2	22.1	17.4
CNN(2conv)	59.2	54.0	22.5	18.1
LSTM	73.0	38.1	19.1	14.1
BiLSTM	**73.9**	41.3	23.2	**21.8**
HA-GRU	72.5	43.5	**23.7**	19.8

Table 3: Multi-label performance: accuracy and *F1* score on two settings for the models when trained on the UMCU discharge letters.

	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

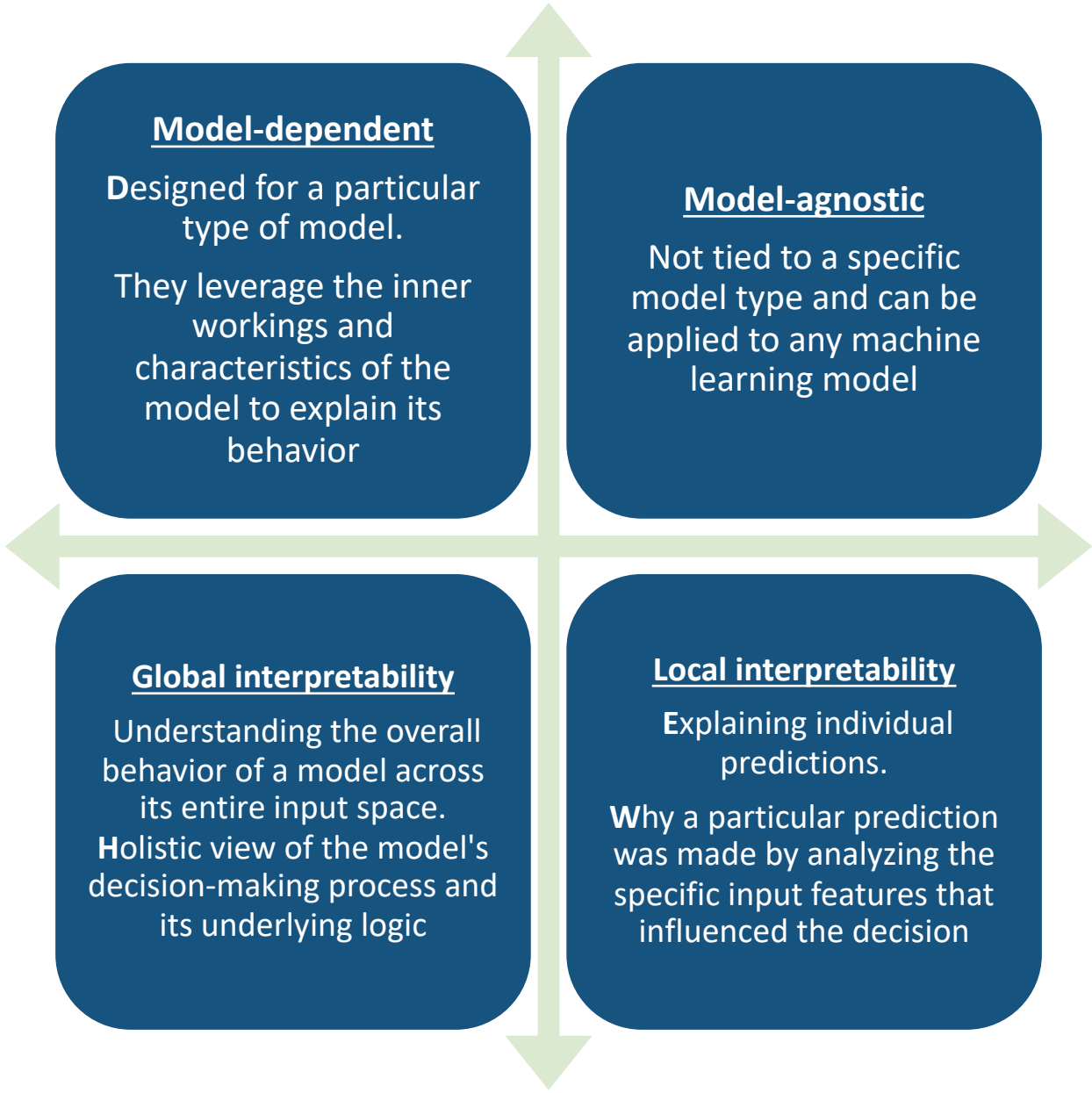
How to know if your results make sense?

Interpretability in Supervised Learning

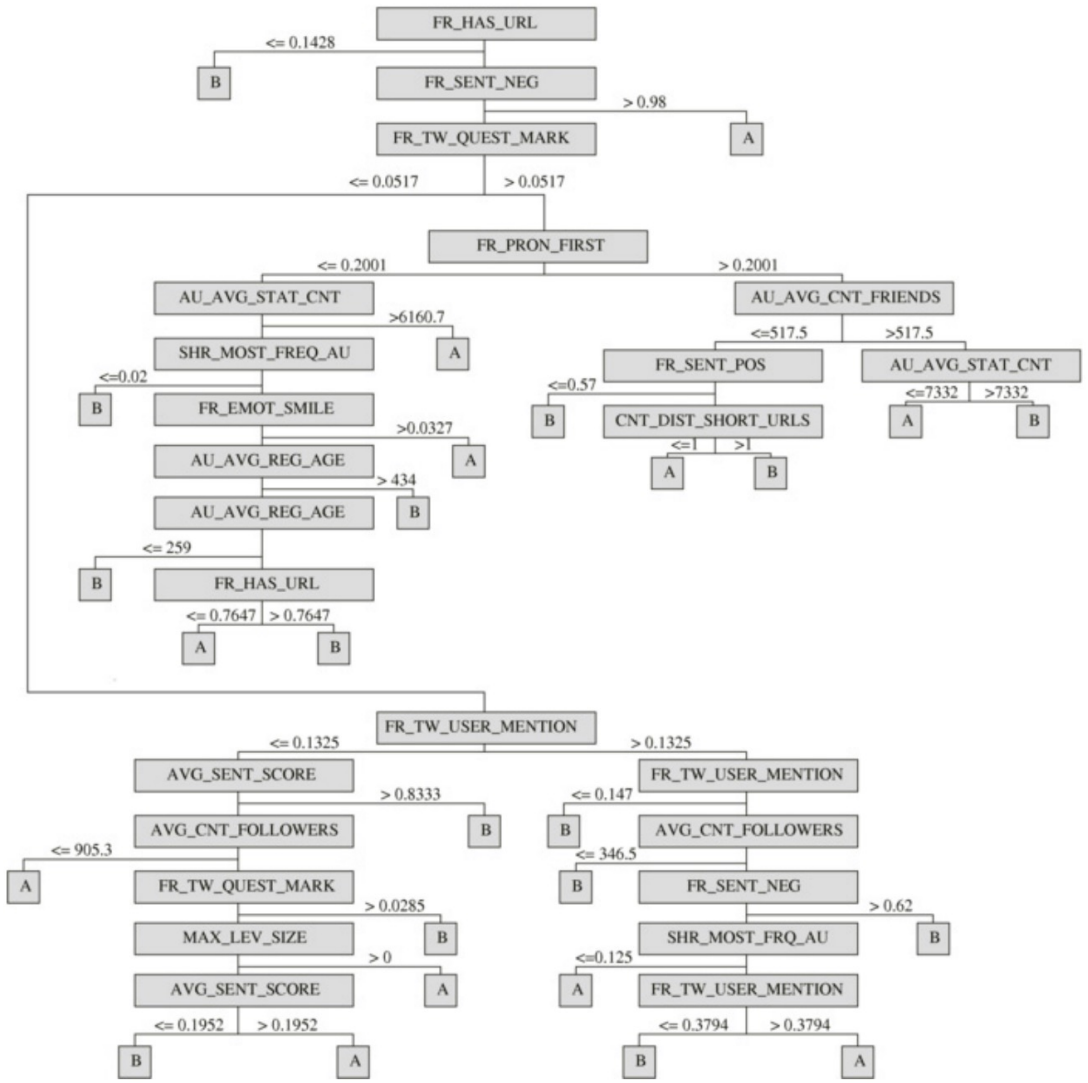


Interpretability

Being right for the right reasons



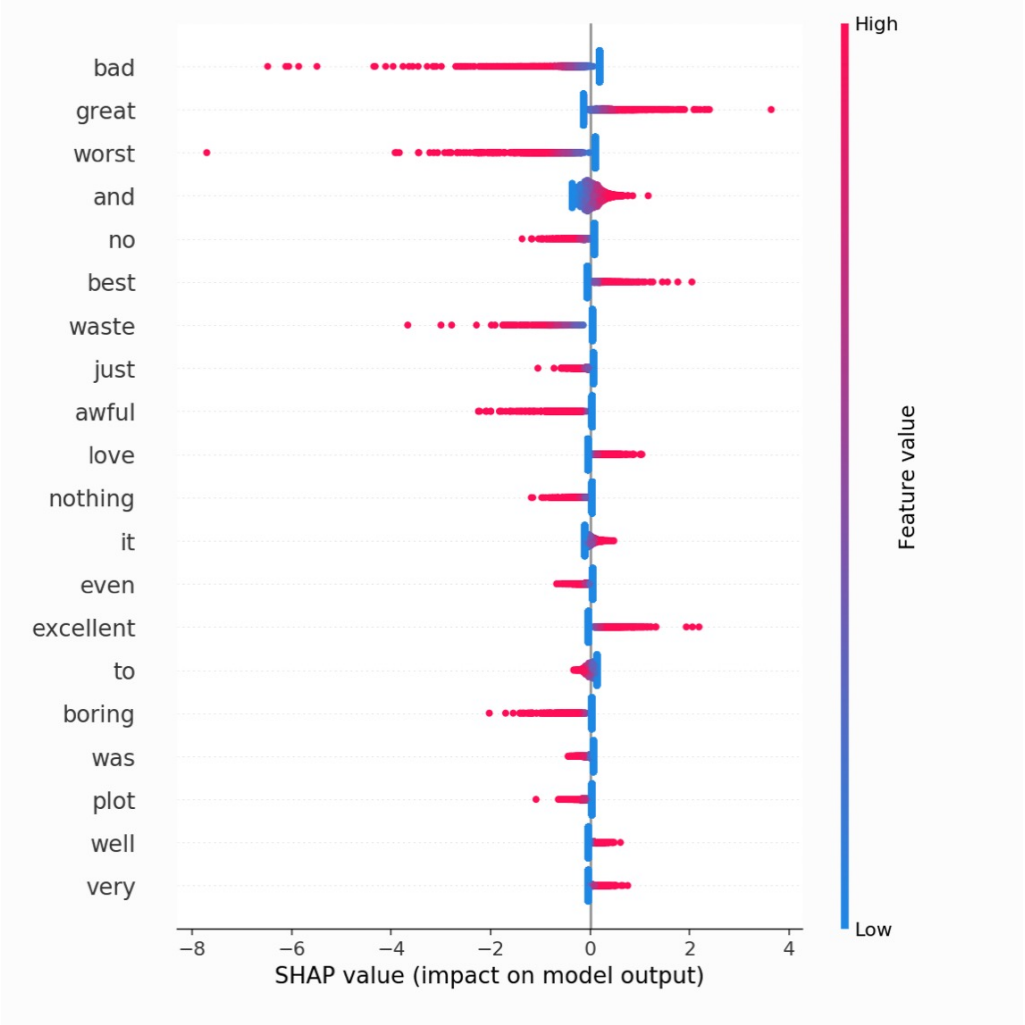
Interpretability: Model-dependent



Global Interpretability

You train a sentiment analysis model

- Analyze the feature importance scores or coefficients of the model
- You find that features related to emotional words have higher importance scores
- This global interpretability analysis reveals the common patterns and factors that contribute to the classification of document as positive or negative



Local Interpretability

You have a trained sentiment analysis model that classifies a document as positive or negative

- You select a specific review classified as positive
  - Why the model made that prediction?
- Analyze the most influential features or words in the article that contributed to the positive classification
- Presence of words like ”good," ”amazing," had a strong positive influence on the model's decision

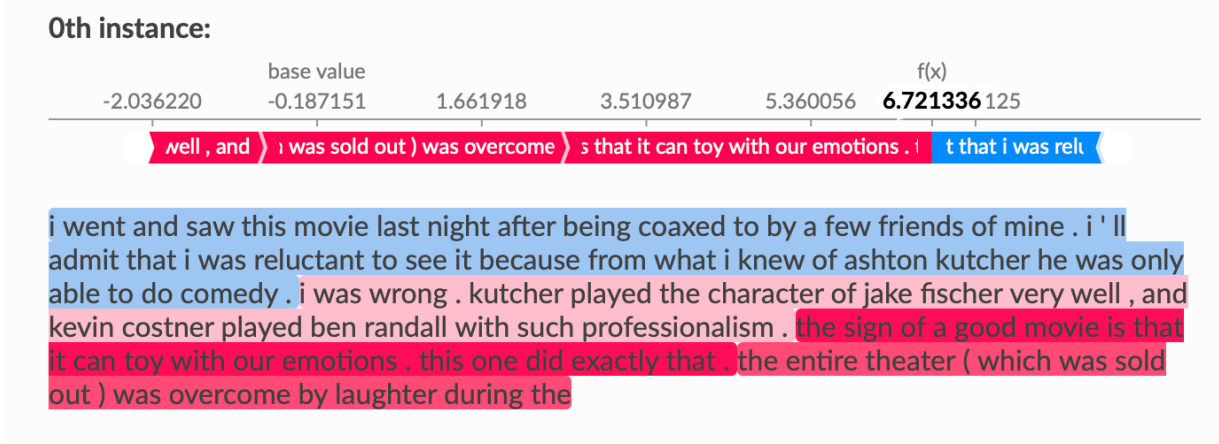
Local interpretability - SHAP

- **SHAP** (SHapley Additive exPlanations) considers different combinations of players (features) to measure their individual contributions.
- It evaluates the prediction for each combination of players and compares it to the prediction when some players are excluded. This helps quantify the importance of each player based on its marginal contribution

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu



https://shap.readthedocs.io/en/latest/example_notebooks/text_examples/sentiment_analysis/Positive%20vs.%20Negative%20Sentiment%20Classification.html

Local interpretability - LIME

“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

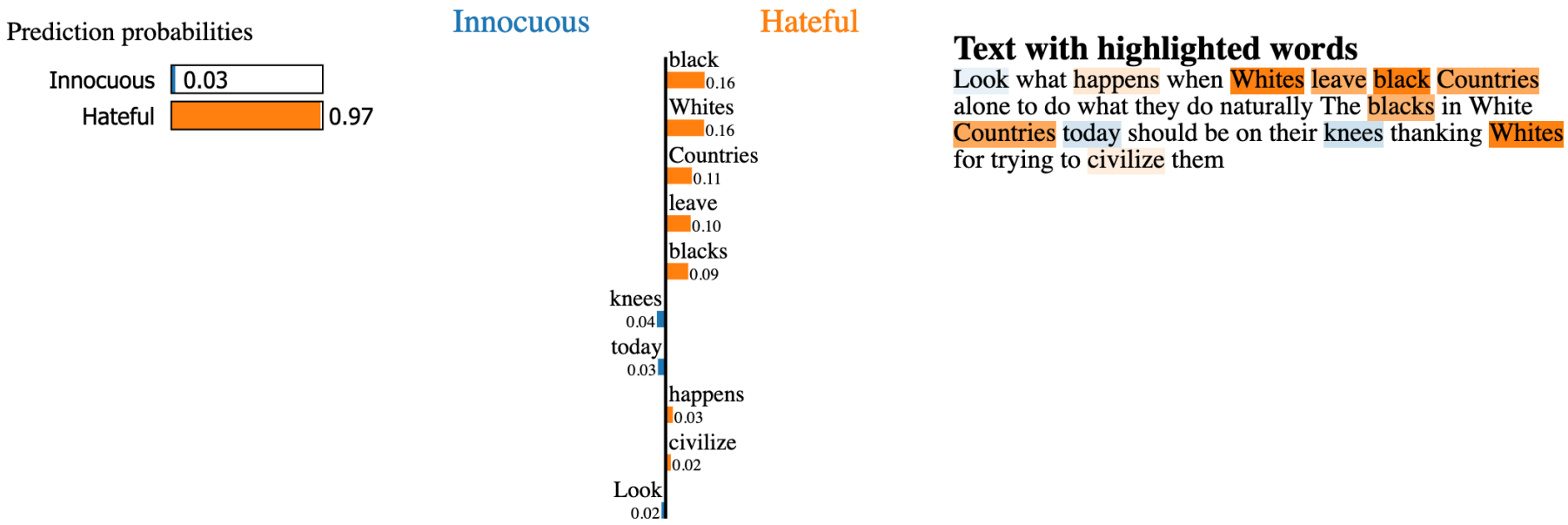
Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

LIME (Local Interpretable Model-Agnostic Explanations) creates simple and interpretable surrogate models for a prediction

It perturbs the input features of an instance and observes how the model's predictions change, allowing to identify the most important features influencing the outcome in a local and understandable way.



Practical 9