Applications of text mining and NLP

Pablo Mosteiro

(with slides by Anastasia Giachanou and Javier García Bernardo)

Summary: the text-mining toolkit

- One-minute paper

Summary: the text-mining toolkit

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

A collection of text mining applications

- Can you think of some text mining applications?
- Think-Pair-Share

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

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 medical reports

Predict stock market returns

Today

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

Today

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

Definition of fake news

- A news article that is intentionally and verifiably false
  - emphasizes both news authenticity and intentions
  - 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: accuracy 55%-58% [0]
- Individuals trust fake news:
  - after repeated exposures (validity effect [1]),
  - if it confirms their preexisting beliefs (confirmation bias [2]),
  - if it pleases them (desirability bias [3])
- Peer pressure “controls” our perception and behavior (e.g., bandwagon effect [4])

[0] Rubin, V. L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. Proceedings of the American Society for Information Science and Technology, 47(1), 1-10.
[1] Boehm, L. E. (1994). The validity effect: A search for mediating variables. Personality and Social Psychology Bulletin, 20(3), 285-293.
[2] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.
[3] Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. Journal of consumer research, 20(2), 303-315.
[4] Leibenstein, H. (1950). Bandwagon, snob, and Veblen effects in the theory of consumers' demand. The quarterly journal of economics, 64(2), 183-207.

Travel fast and more

- Compared to the truth, fake news on Twitter
  - is retweeted by many more users, and
  - spreads far more rapidly (especially political)
- During the 2016 U.S. presidential election campaign:
  - top 20 fake election stories generated **8,711,000** shares, reactions, and comments on Facebook
  - **7,367,000** for the top 20 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

- Where does text mining come in?

The role of content

- Fake news ?= truth:
  - writing style and quality (Undeutsch hypothesis)
  - quantity such as word counts (information manipulation theory)
  - sentiments expressed (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

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

Information credibility on Twitter

- Assessing the credibility of a given set of tweets
- Data collection: all tweets matching queries in 2-day window; 2500 topics
- Features:
  - Message-based: length, presence of special chars, sentiment, etc.
  - User-based: age, number of followers, number followed, etc.
  - Topic-based: fraction of tweets with URL, fraction of positive, etc.
  - Propagation-based: depth of 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

Information credibility on Twitter

- Assessing the credibility of a given set of tweets
- Data collection: all tweets matching queries in 2-day window; 2500 topics
- Features:
  - Message-based: length, presence of special chars, sentiment, etc.
  - User-based: age, number of followers, number followed, etc.
  - Topic-based: fraction of tweets with URL, fraction of positive, etc.
  - Propagation-based: depth of re-tweet, number of initial tweets of a topic
- Decision Tree classifier
  - Evaluation?

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

Information credibility on Twitter: Results

F1 = 0.7-0.8

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 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

Today

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

Hate Speech

- Social media enable its propagation
- Hate speech detection crucial to reducing crime, protecting people
  - 2023: D66 leader Sigrid Kaag stopped with politics. She mentions "hate, intimidation and threats" and the effect on her family as the reason to stop [0]

[0] https://nos.nl/nieuwsuur/artikel/2482833-toelichting-twitter-reacties-op-vertrek-sigrid-kaag

Hate Speech Definition

- 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**’.

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

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?

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.

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

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.

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.

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

Today

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

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

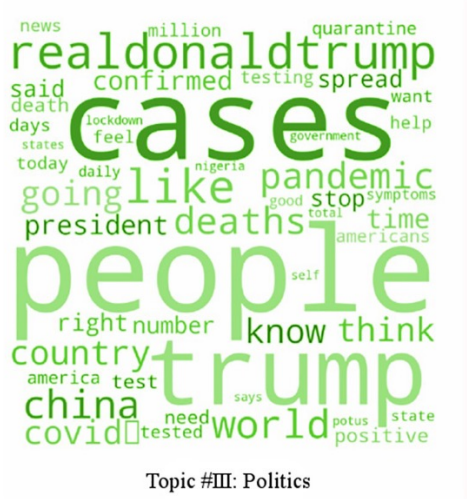
Method: LDA (4 and 16 topics) + correlation of topics with frames

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

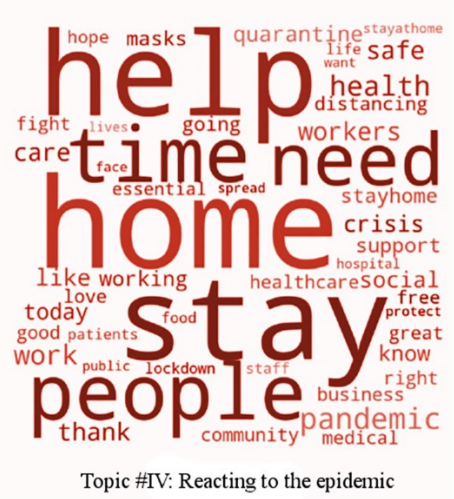
COVID pandemic (Wicke & Bolognesi 2020)

Topic #I: Communications and Reporting

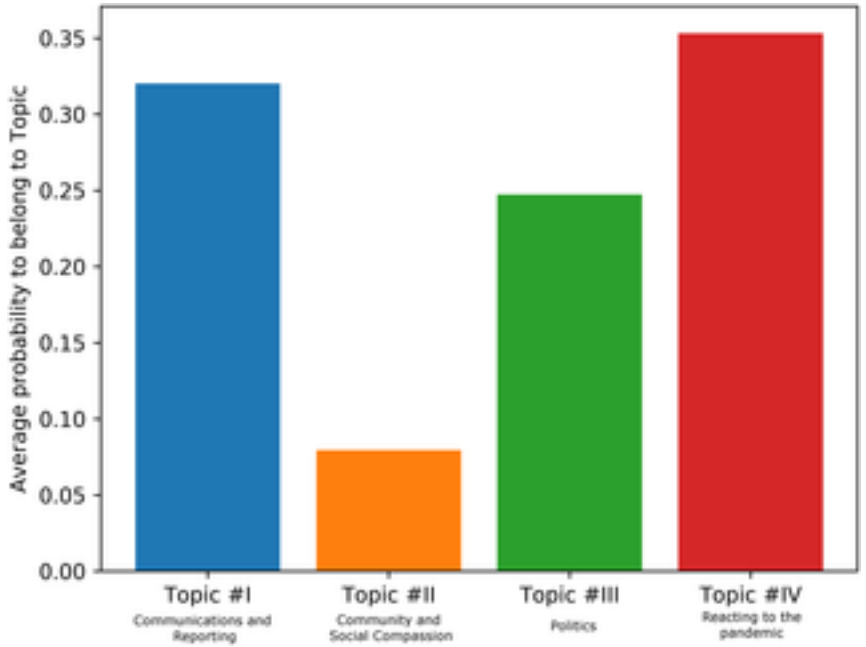
Topic #II: Community and Social Compassion



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

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** ▾

Today

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

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

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 discharge 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:

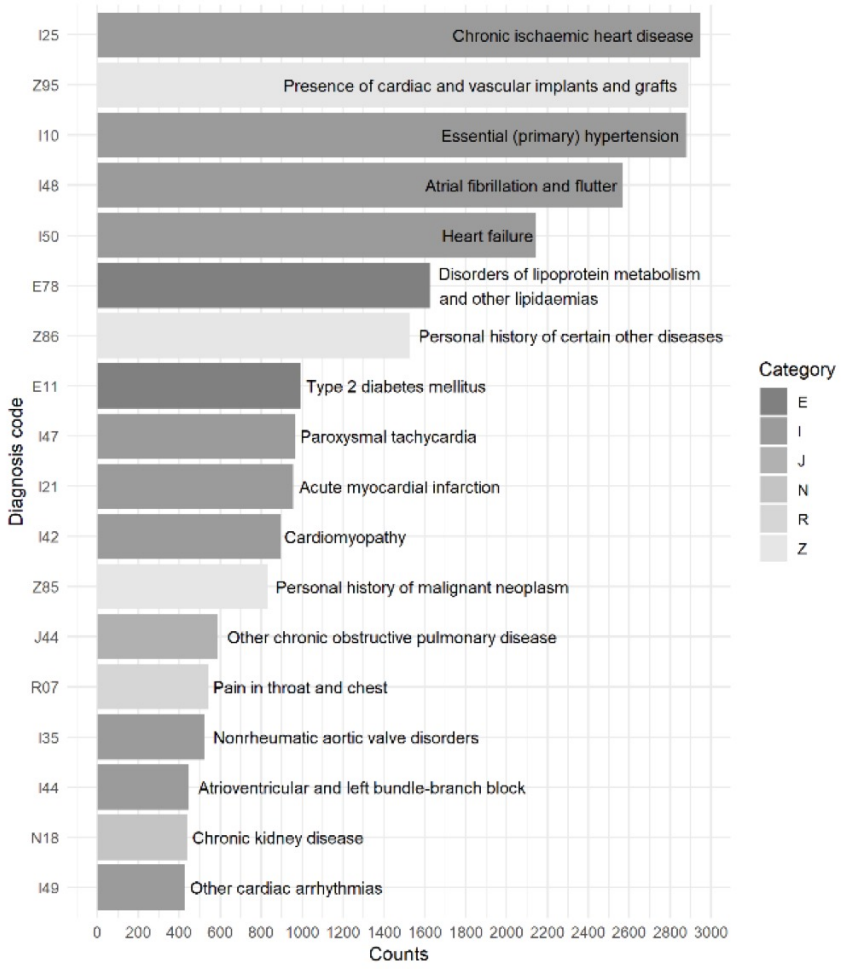


Figure 1: ICD rolled-up codes with more than 400 appearances in the UMCU dataset.

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

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

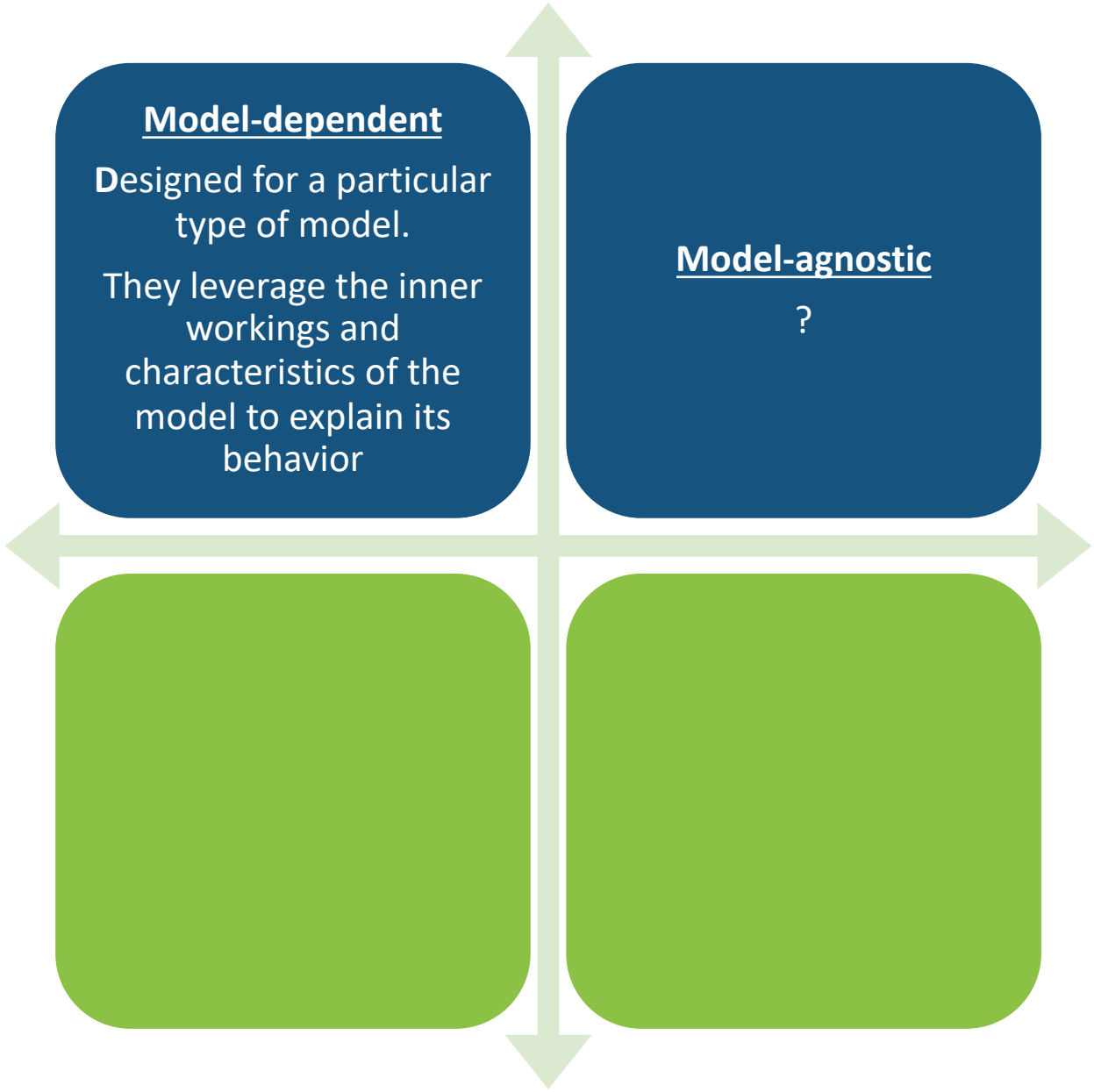
Arjan Sammani, Ayoub Bagheri, Peter G. M. van der Heijden, Anneline S. J. M. te Riele, Annette F. Baas, C. A. J. Oosters, Daniel Oberski & Folkert W. Asselbergs

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

How to know if your results make sense?

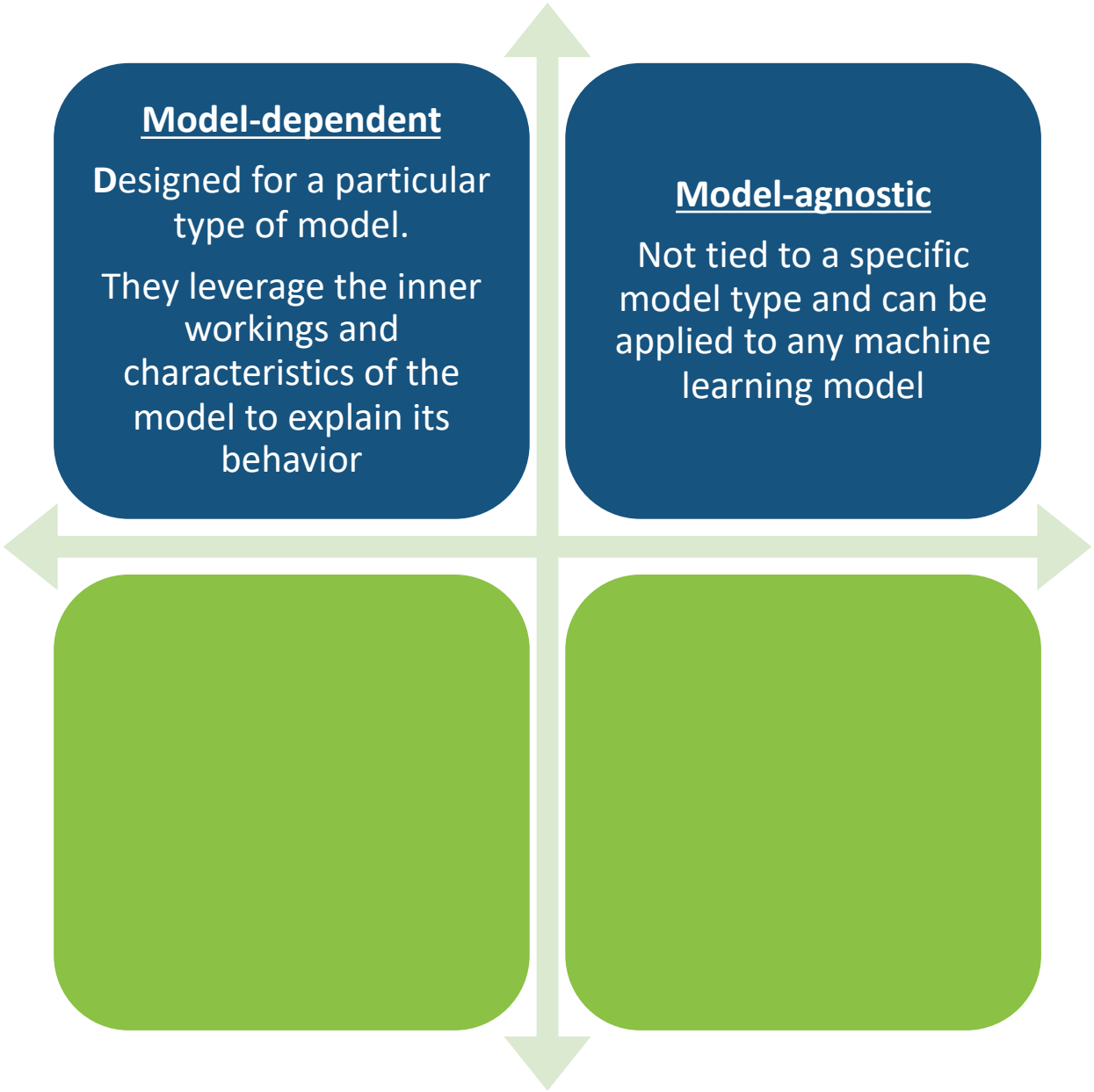
Interpretability

Being right for the right reasons



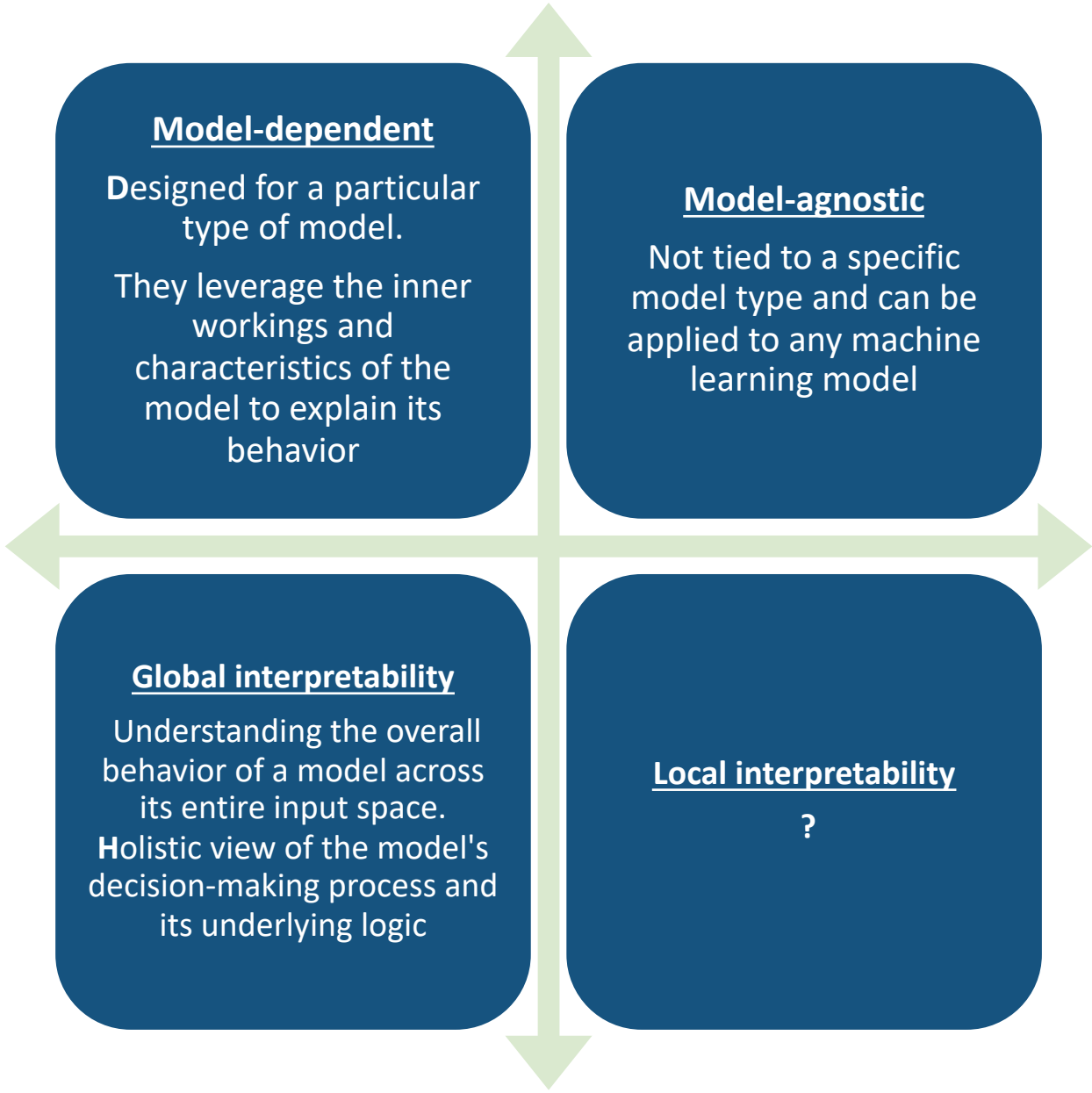
Interpretability

Being right for the right reasons



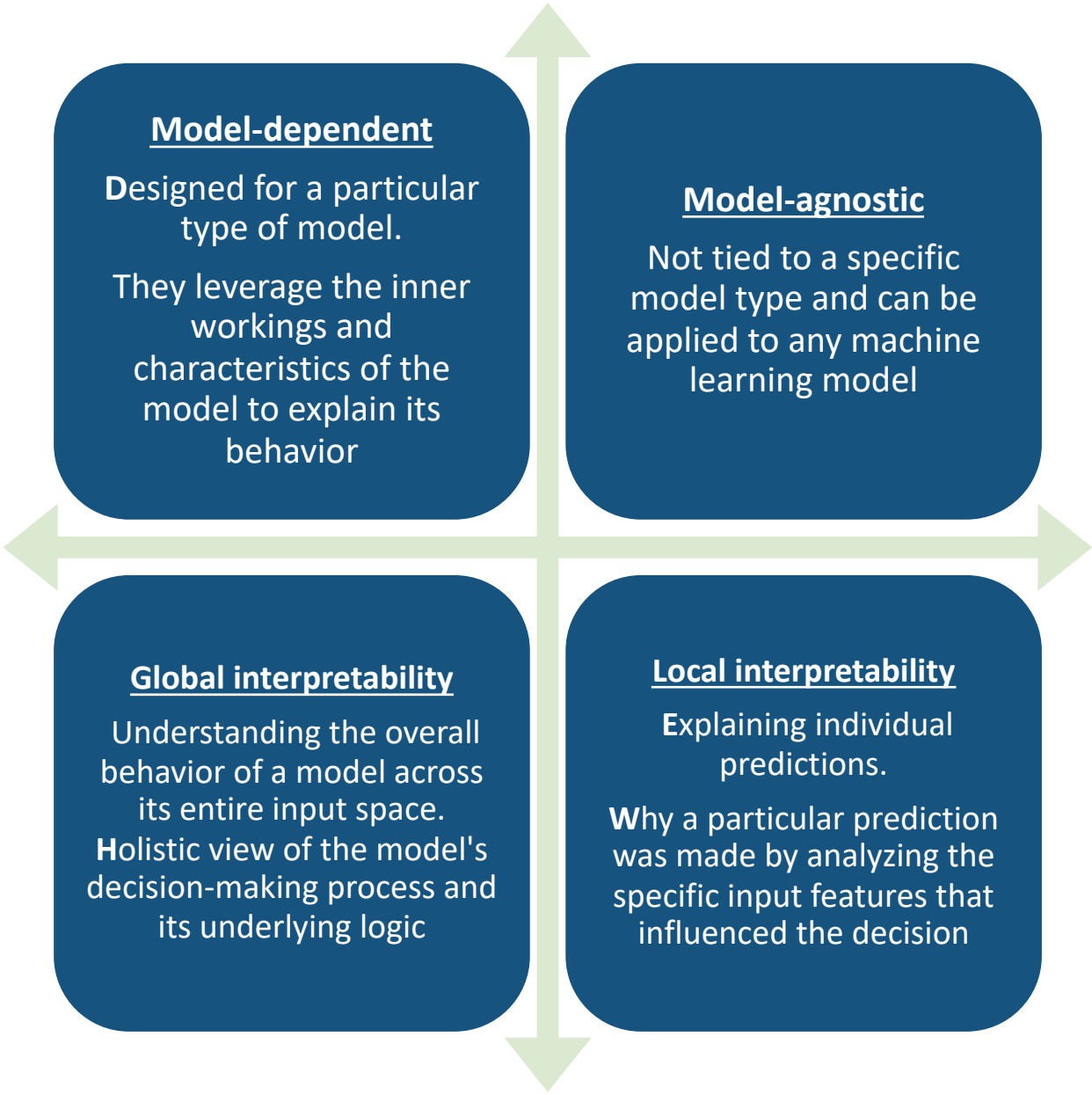
Interpretability

Being right for the right reasons

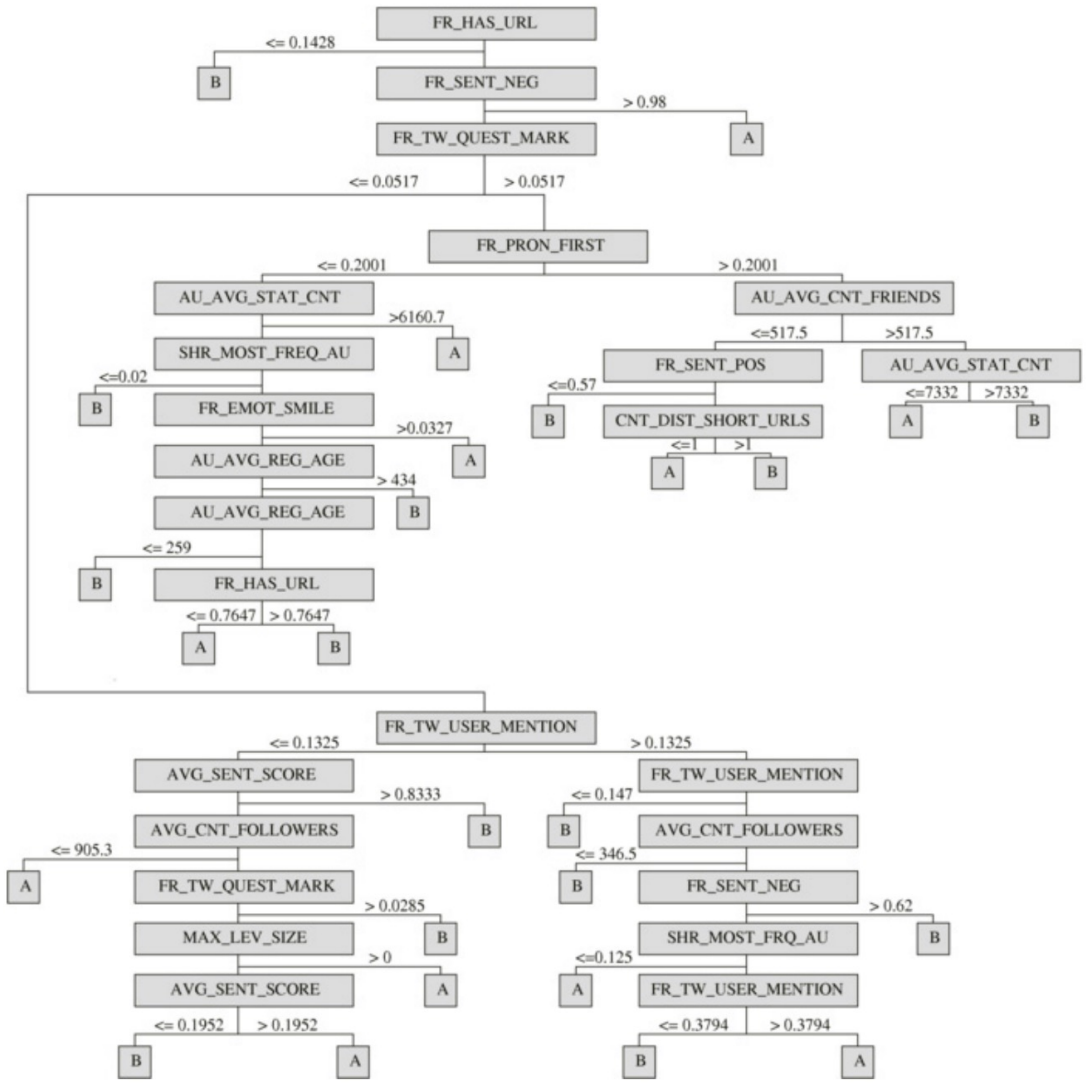


Interpretability

Being right for the right reasons



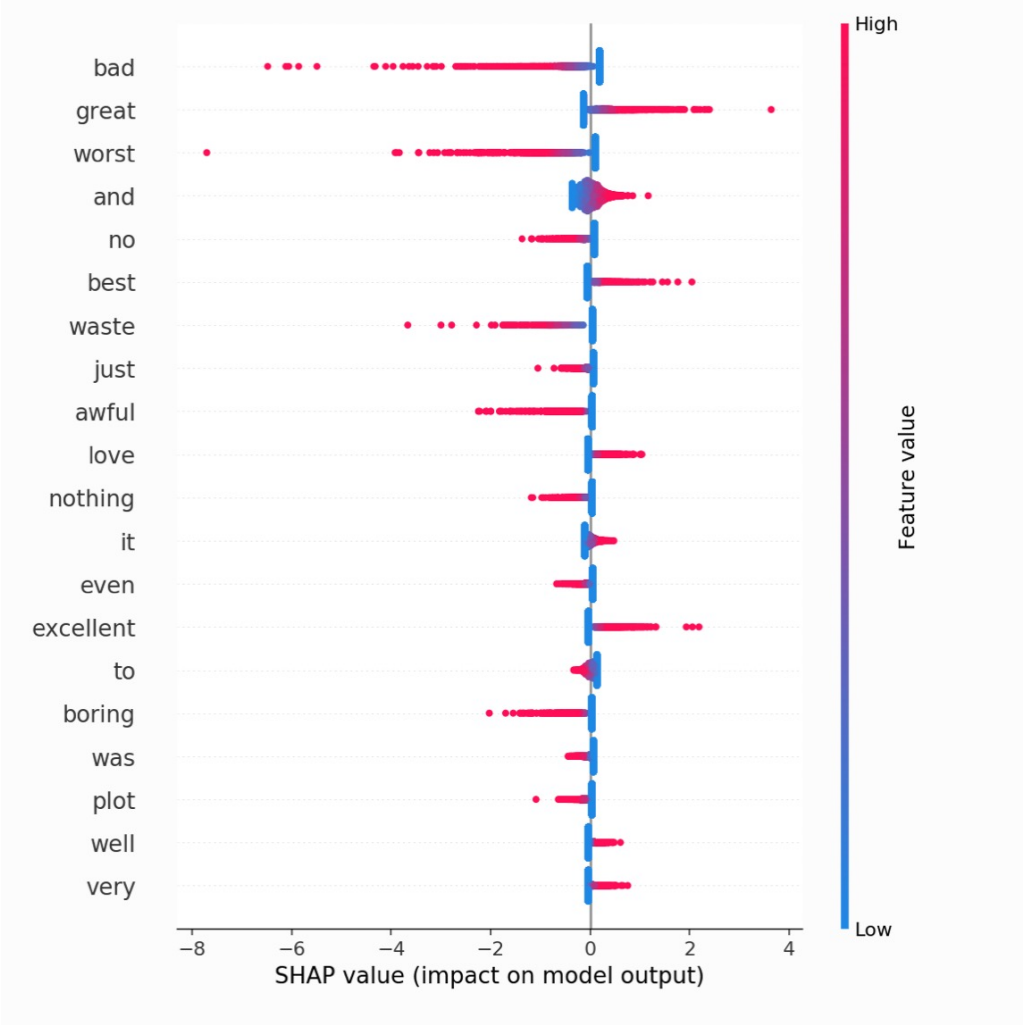
Interpretability: Model-dependent



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 - LIME

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

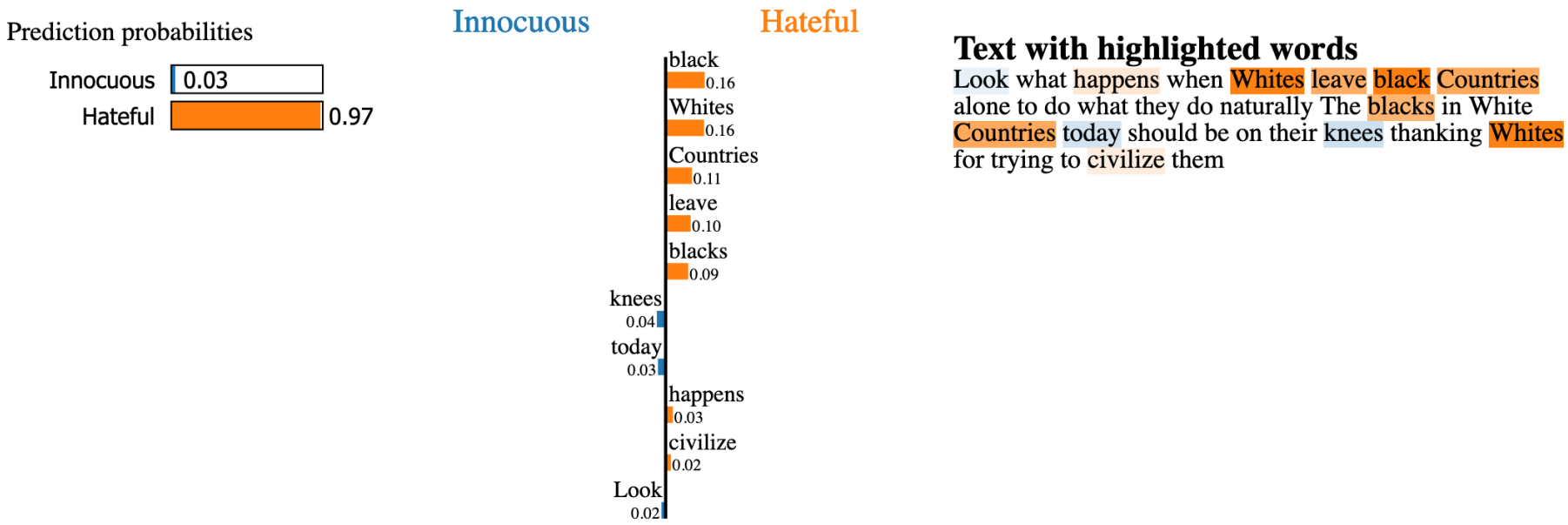
Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

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