

Responsible Text Mining and Applications

**Text Mining, Transforming Text into
Knowledge**

*Anastasia Giachanou,
Ayoub Bagheri*



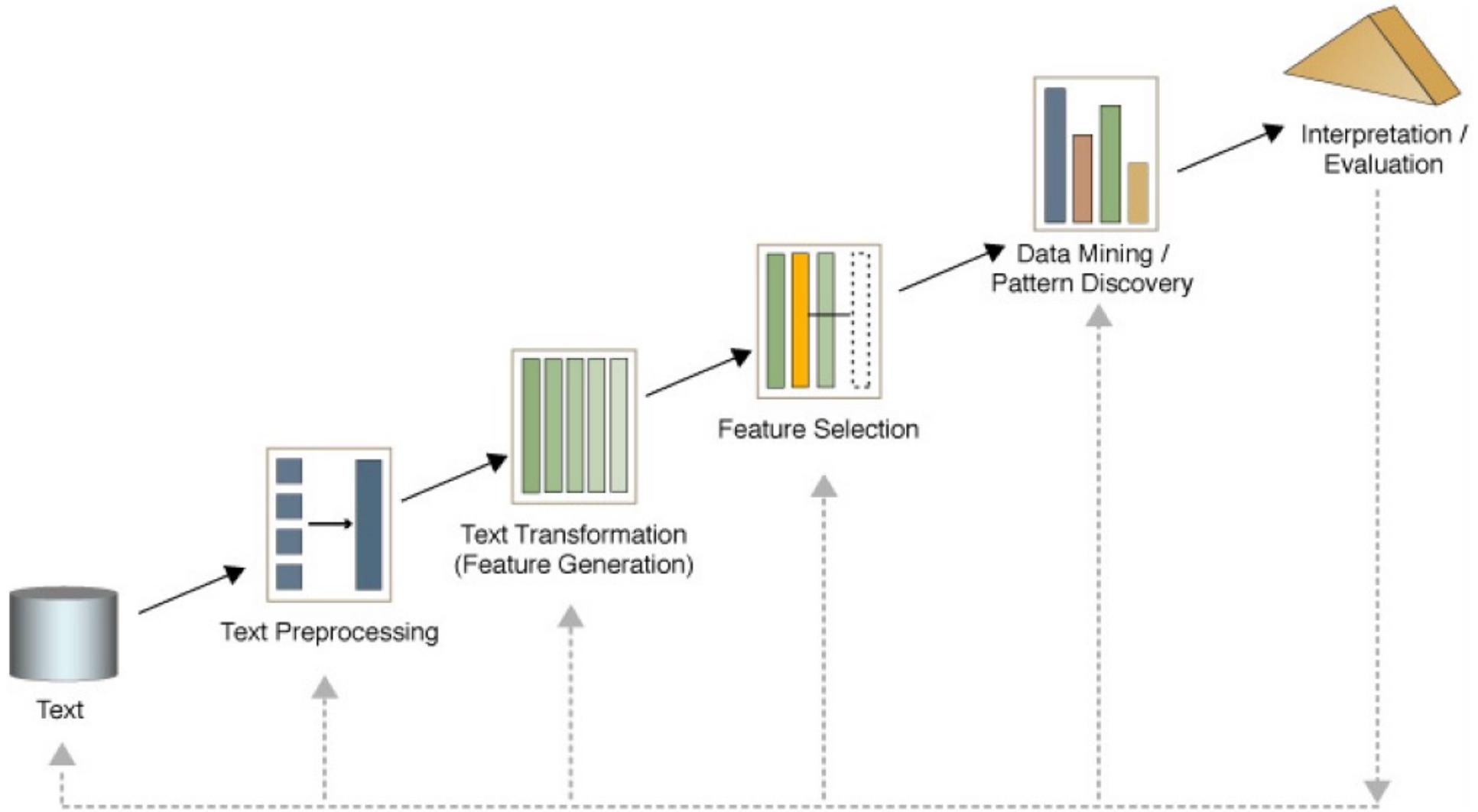
Last week

- Sentiment analysis
 - Lexicon and dictionary-based sentiment analysis
 - Machine learning-based sentiment analysis
- Sentiment classification
 - Document level
 - Sentence level

Today

- Responsible text mining
- More applications
- Interpretability in NLP
- Future of text mining & NLP

Text mining process



Text mining process

- **Data: Text**
- **Text Preprocessing:** is the process of cleaning, normalizing, and structuring raw text data into a format suitable for analysis or input into NLP models. (week 2)
- **Text transformation, feature generation:** involves converting text data into a different format or structure, such as numerical vectors or simplified forms, to make it suitable for analysis or modeling. (weeks 1, 2, 3, 6, 7, 8)
- **Feature selection:** is the process of identifying and selecting the most relevant features from a dataset to improve model performance and reduce complexity. (week 4)
- **Data mining, pattern discovery:** is the process of extracting meaningful patterns and knowledge from text. (weeks 3, 5, 7, 8, 9)
- **Interpretation / Evaluation:** is the process of understanding and explaining the model and patterns / is the assessment process to measure performance and quality. (weeks 3-9)

Responsible Text Mining

Dual use: Should I build this system?

Predicting Depression via Social Media

Munmun De Choudhury

Michael Gamon

Scott Counts

Eric Horvitz

Microsoft Research, Redmond WA 98052

{munmund, mgamon, counts, horvitz}@microsoft.com

“We explore the potential to use social media to detect and diagnose major depressive disorder in individuals.”

How can such a system be used for a beneficial purpose?
How can such a system be used for a harmful purpose?

(3 min)

Responsible text mining

Are we really
measuring what we
intend to measure?





October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

<https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html>

POLICY

Predictive policing is still racist—whatever data it uses

Training algorithms on crime reports from victims rather than arrest data is said to make predictive tools less biased. It doesn't look like it does.

By Will Douglas Heaven

February 5, 2021

Gender

can you complete the sentence 10 times "the secretary was typing when"

Sure! Here are 10 different ways to complete the sentence:

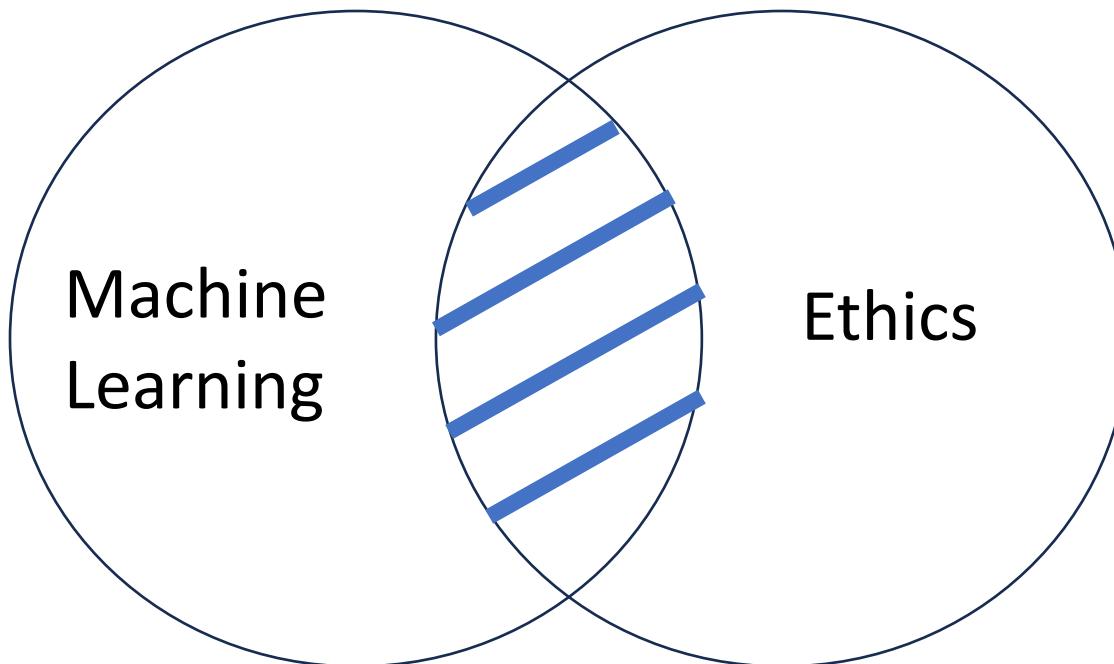
1. **The secretary was typing when** the phone suddenly rang.
2. **The secretary was typing when** the power went out.
3. **The secretary was typing when** her boss called her into the office.
4. **The secretary was typing when** she realized she had made a mistake in the report.
5. **The secretary was typing when** a loud noise outside distracted her.
6. **The secretary was typing when** her computer screen froze.
7. **The secretary was typing when** a colleague asked her for help.
8. **The secretary was typing when** she accidentally spilled coffee on her keyboard.
9. **The secretary was typing when** the fire alarm started ringing.
10. **The secretary was typing when** she received an urgent email.

Would you like more variations? 😊



Do you think that these results are biased?
If no, why not?
If yes, where this bias comes from? Should
openAI try to fix it?

What is algorithmic fairness



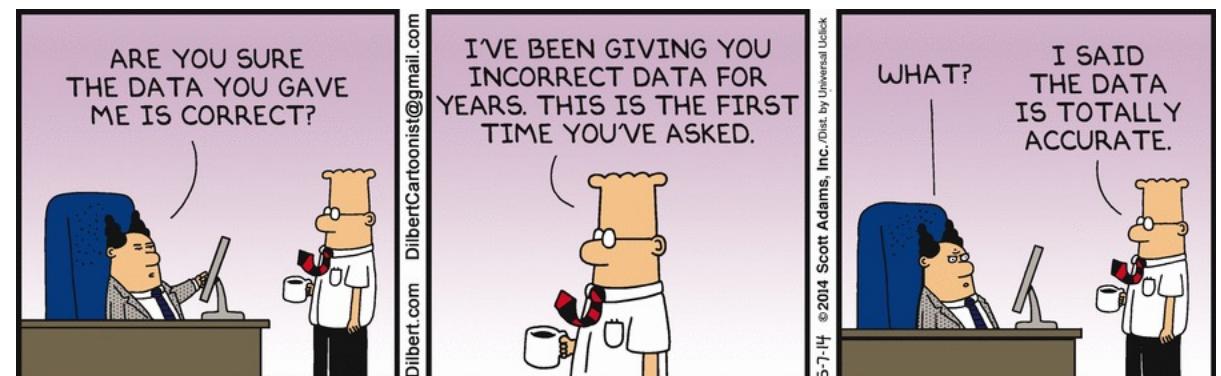
- Causes of bias in data and algorithms
- Measurements of fairness
- Creating fair algorithms
- Regulate machine learning

What is bias?

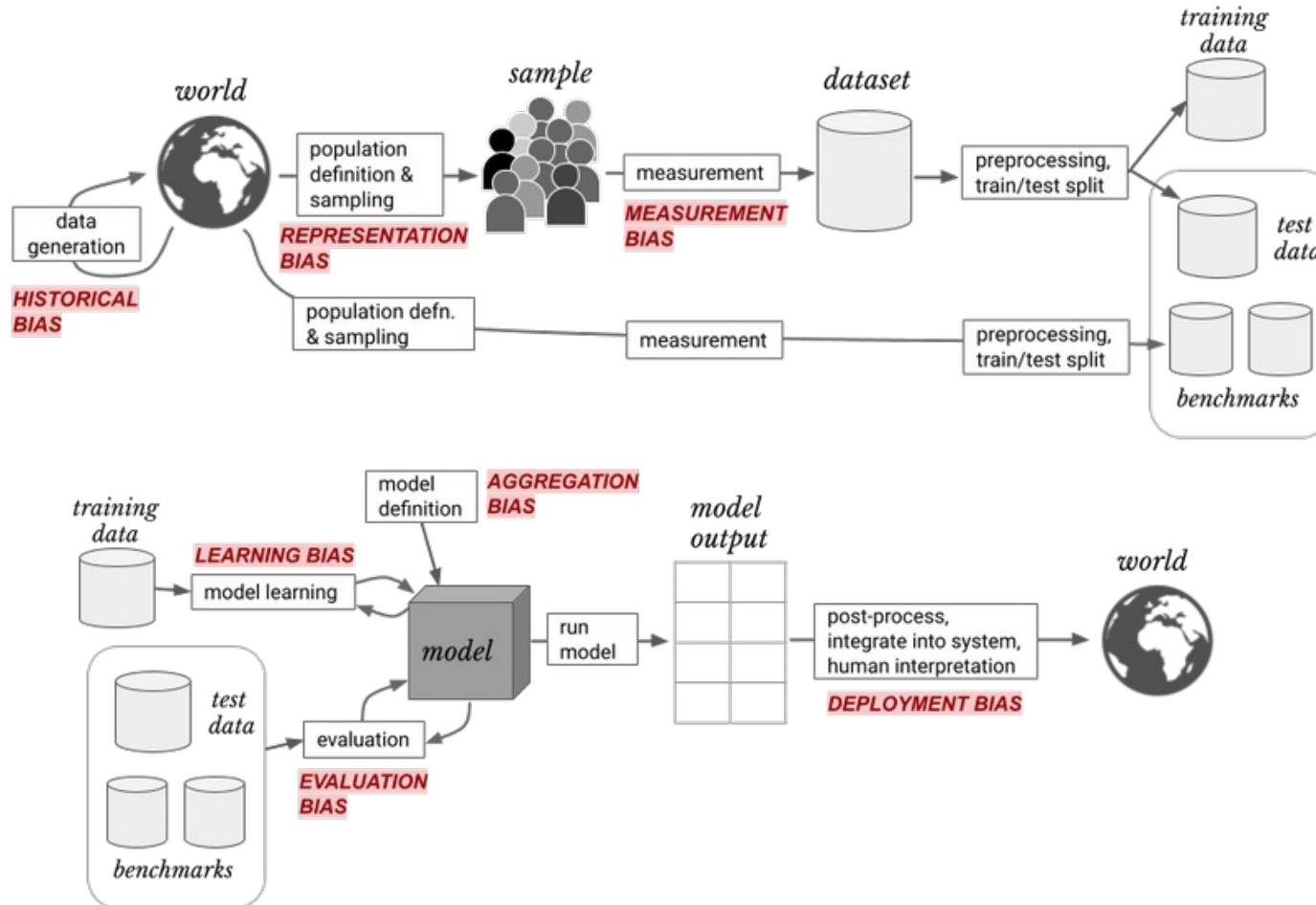
- Prejudice towards or against one person or group
- There is no single definition of bias and fairness in the literature
- Any systematic error that can result in an *unfair* model, a model that is prejudiced towards or against one person or group

Sources of Bias - Data

- No real-world datasets that are free of societal biases



Sources of bias



Historical bias

- The world *as it is* or *was* leads to a model that produces harmful outcomes

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Representation bias

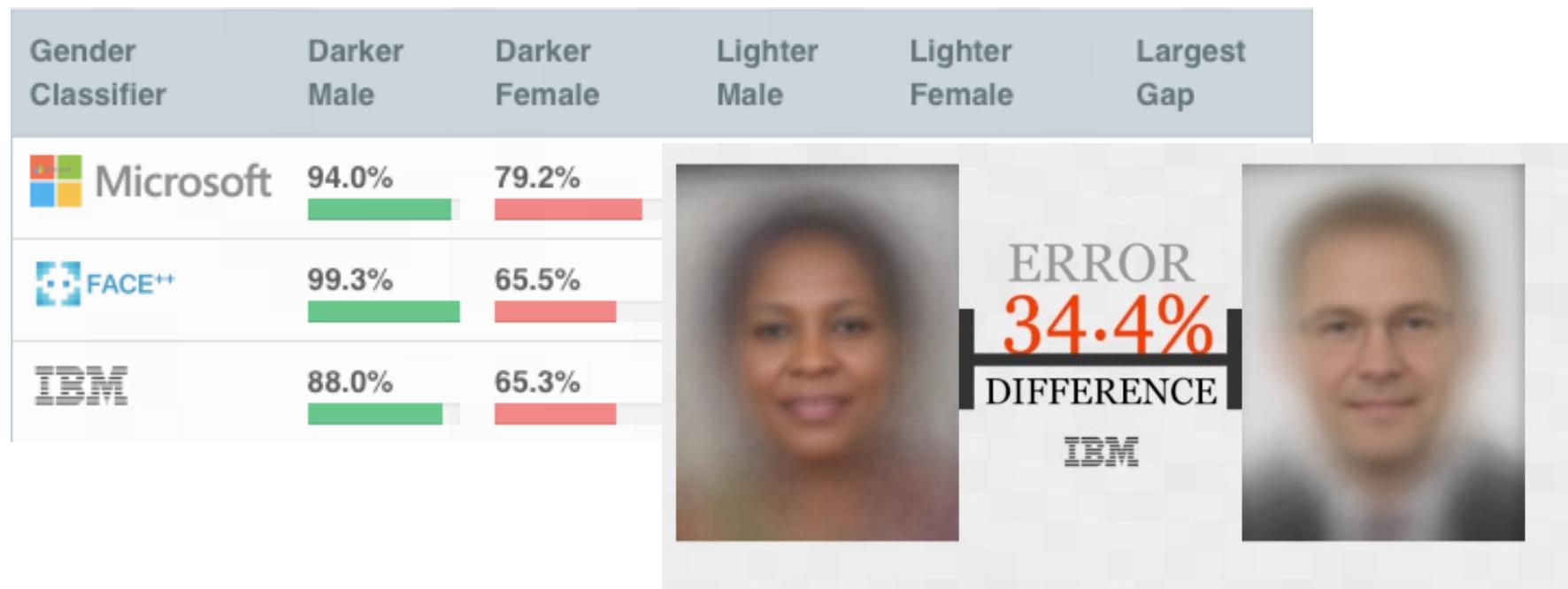
- The development sample underrepresents some part of the population

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT* '18), 77–91.

Representation bias

- The development sample underrepresents some part of the population



Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT* '18), 77–91.

Feedback loop

- You build a system to filter out CVs for technical jobs
- You end up hiring more men for those jobs since many women are ruled out
- You feed the new data to the system
- Bias is amplified

Fairness via awareness

- Okay then let's remove them

Fairness via awareness

- Okay then let's remove sensitive features (e.g., gender pronouns, race)
- The remaining features may correlate with the sensitive features
- Very common in the high-dimensional data in modern ML
- Proxies (zip code correlated with race)
 - In effect, Amazon's system taught itself that male candidates were preferable. It penalized résumés that included the word "women's", as in "women's chess club captain". And it downgraded graduates of two all-women's colleges, according to people familiar with the matter.
- Or the Amazon case:

Fairness metrics

- Fairness metrics
 - Demographic Parity
 - Equalized Odds
 - Disparate Impact
- Compare model performance across subgroups
 - Measure accuracy, precision, recall, and F1-score separately for different demographic groups (e.g., gender, race, age).

Break

Text Mining Applications

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

Some applications

- Fake news detection
- Hate speech detection
- Healthcare applications

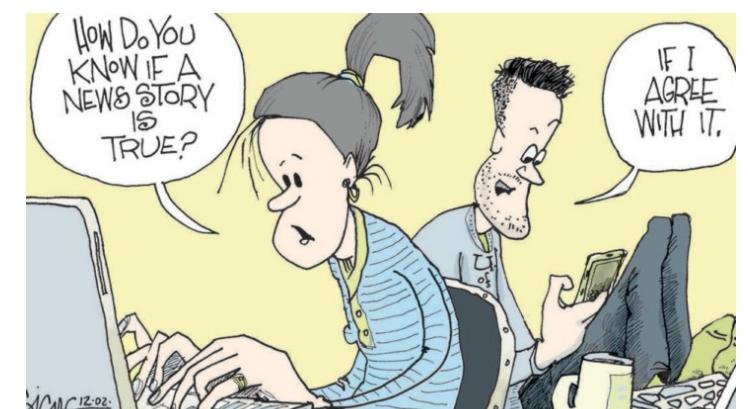
Fake News Detection

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

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 (now X)
 - 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 Castillo¹

Marcelo Mendoza^{2,3}

Barbara Poblete^{2,4}

{chato,bpoblete}@yahoo-inc.com, marcelo.mendoza@usm.cl

¹Yahoo! Research Barcelona, Spain

²Yahoo! Research Latin America, Chile

³Universidad Técnica Federico Santa María, Chile

⁴Department of Computer Science, University of Chile

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

Kashyap Popat¹, Subhabrata Mukherjee², Andrew Yates¹, Gerhard Weikum¹

¹Max Planck Institute for Informatics, Saarbrücken, Germany

²Amazon Inc., Seattle, USA

Fake News Early Detection: A Theory-driven Model

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

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?

Information credibility on Twitter: Results

F1 = 0.7-0.8

Fake news detection

Information Credibility on Twitter

Carlos Castillo¹

Marcelo Mendoza^{2,3}

Barbara Poblete^{2,4}

{chato,bpoblete}@yahoo-inc.com, marcelo.mendoza@usm.cl

¹Yahoo! Research Barcelona, Spain

²Yahoo! Research Latin America, Chile

³Universidad Técnica Federico Santa María, Chile

⁴Department of Computer Science, University of Chile

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

Kashyap Popat¹, Subhabrata Mukherjee², Andrew Yates¹, Gerhard Weikum¹

¹Max Planck Institute for Informatics, Saarbrücken, Germany

²Amazon Inc., Seattle, USA

Fake News Early Detection: A Theory-driven Model

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

Hate Speech Detection

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

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

Zeerak Waseem, James Thorne and Joachim Bingel

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

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?

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

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.

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

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

Zeerak Waseem, James Thorne and Joachim Bingel

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

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

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

Healthcare Applications

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](#)

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:

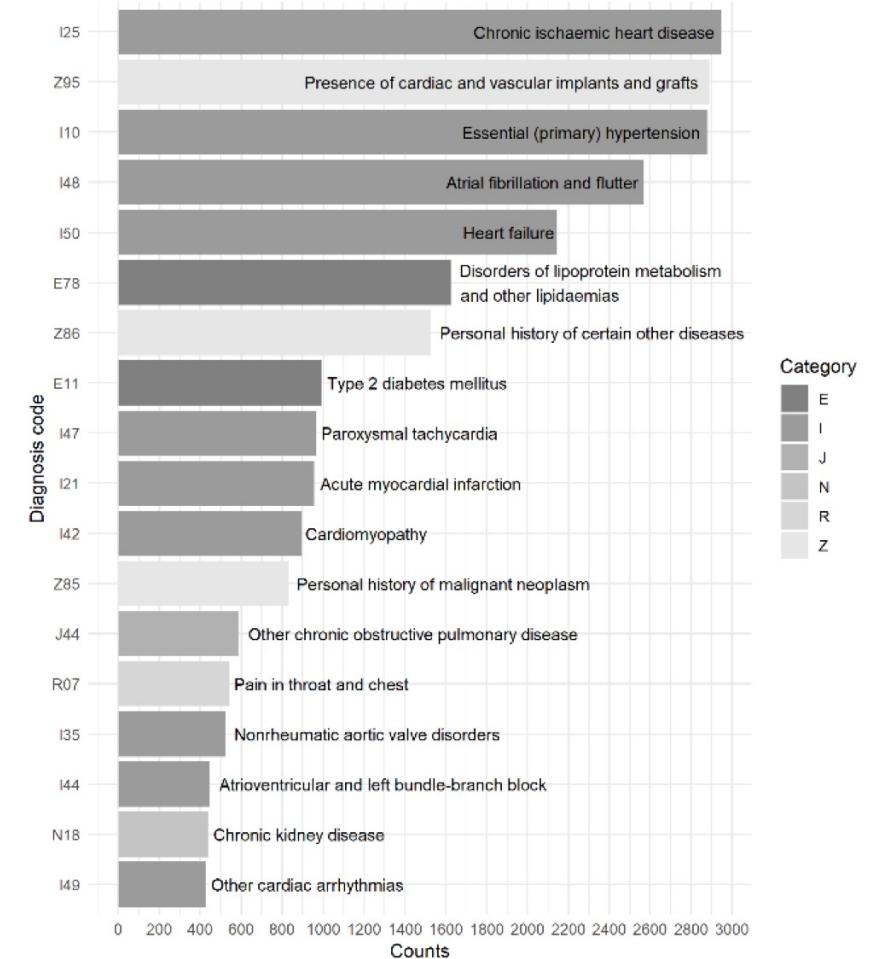


Figure 1: ICD rolled-up codes with more than 400 appearances in the UMCU dataset.

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

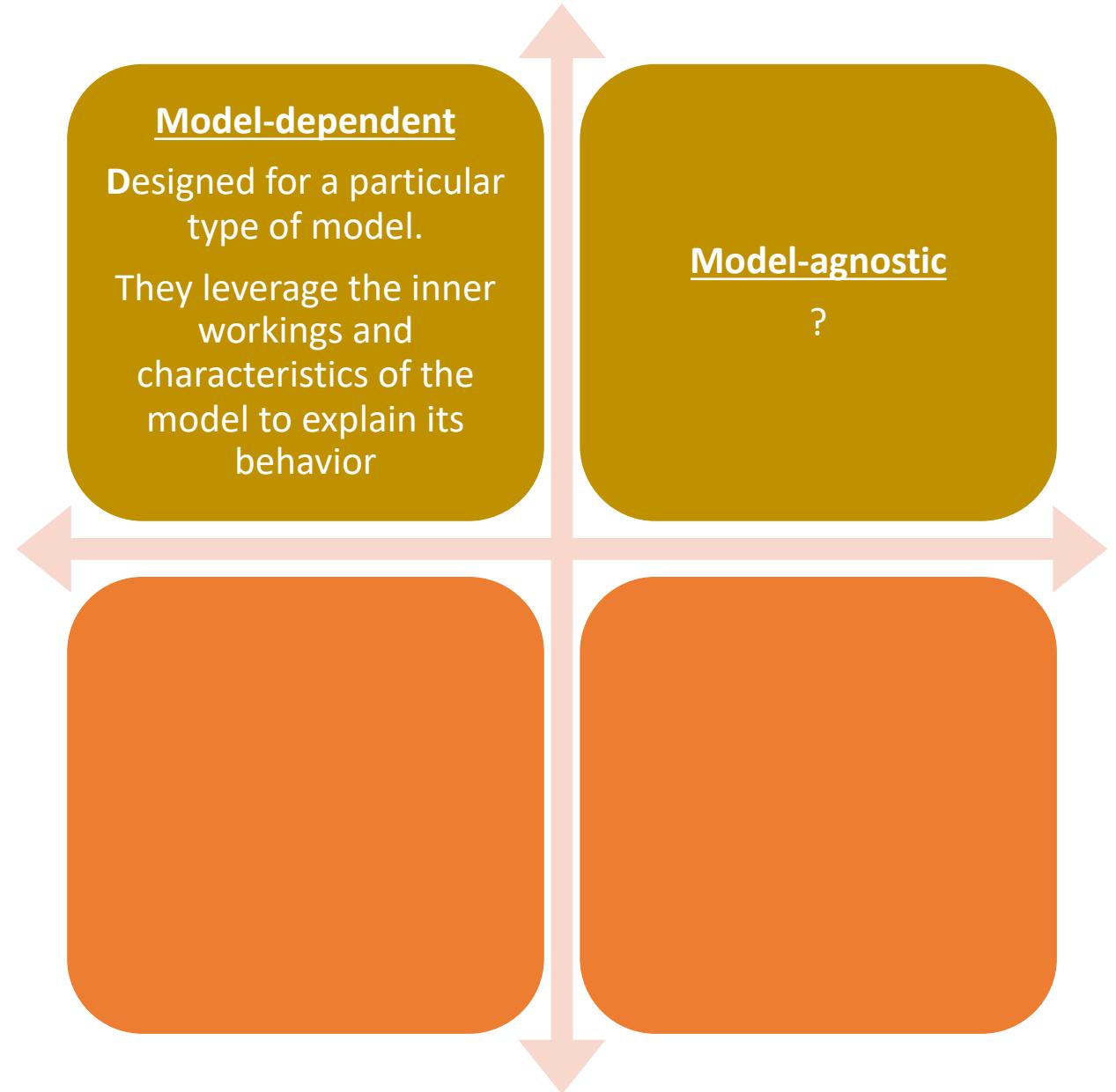
[Arjan Sammani](#) , [Ayoub Bagheri](#), [Peter G. M. van der Heijden](#), [Anneline S. J. M. te Riele](#), [Annette F. Baas](#), [C. A. J. Oosters](#), [Daniel Oberski](#) & [Folkert W. Asselbergs](#)

Interpretability in NLP

How to know if your results make sense?

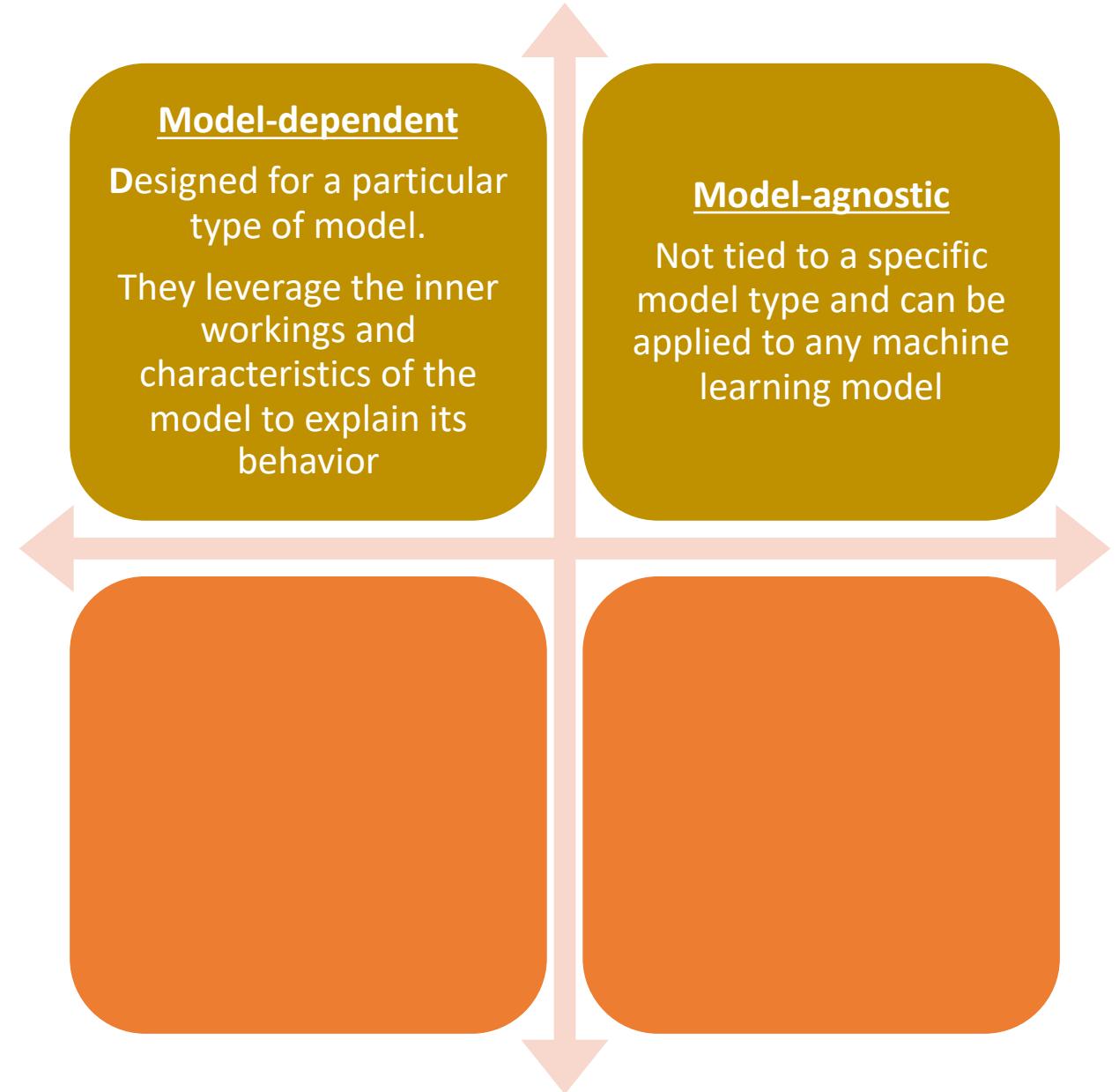
Interpretability

Being right for the right reasons



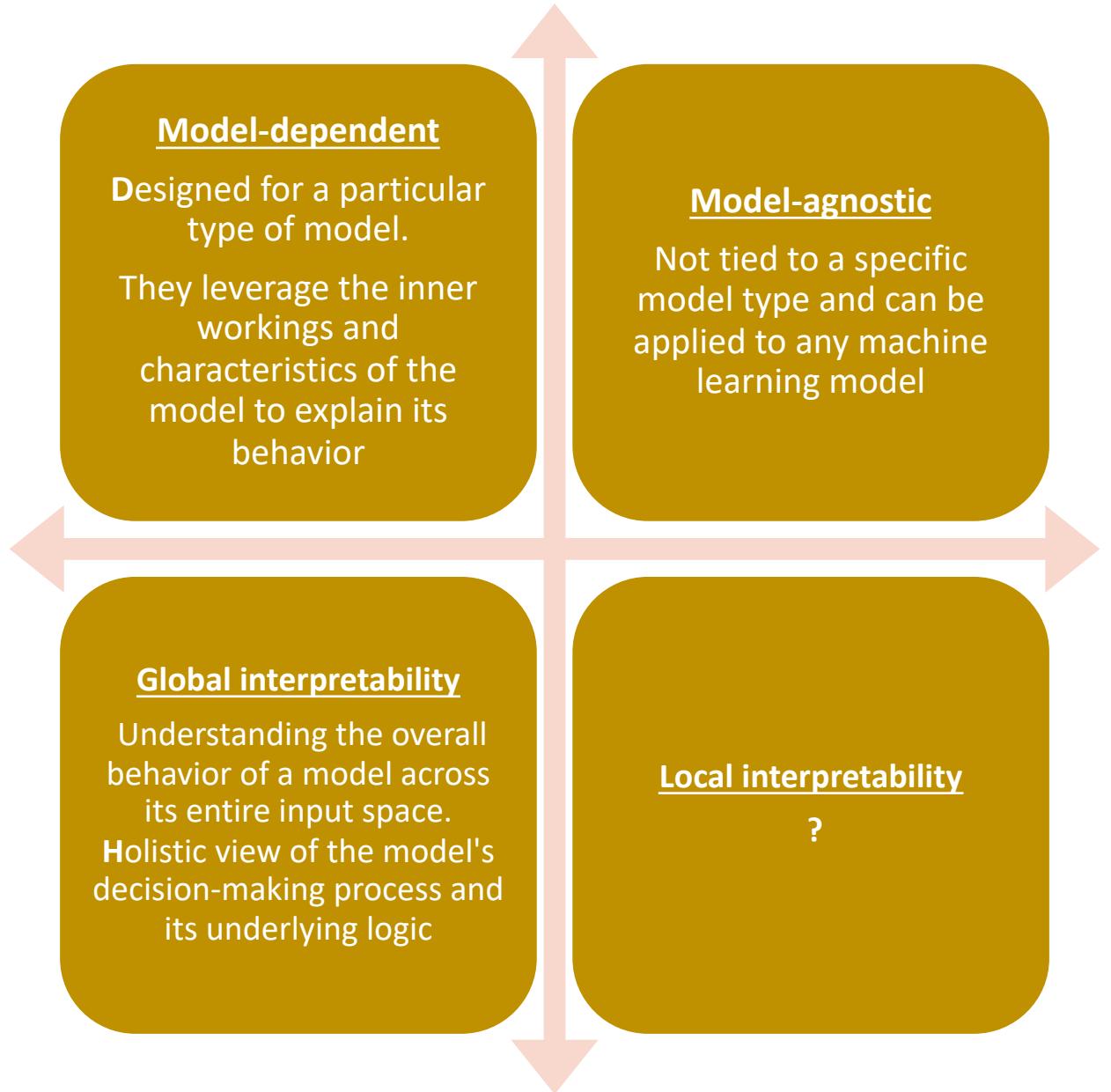
Interpretability

Being right for the right reasons



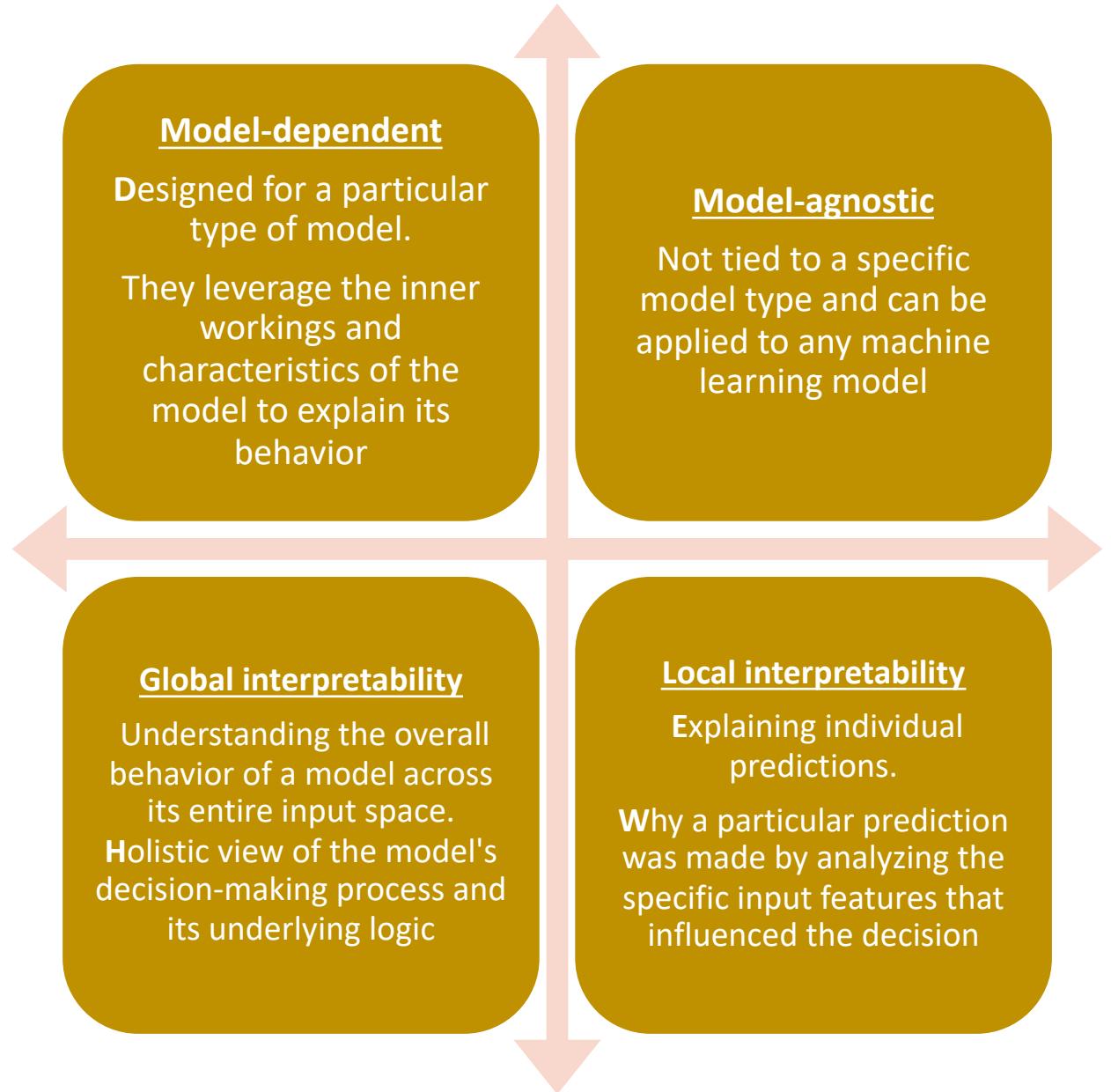
Interpretability

Being right for the right reasons

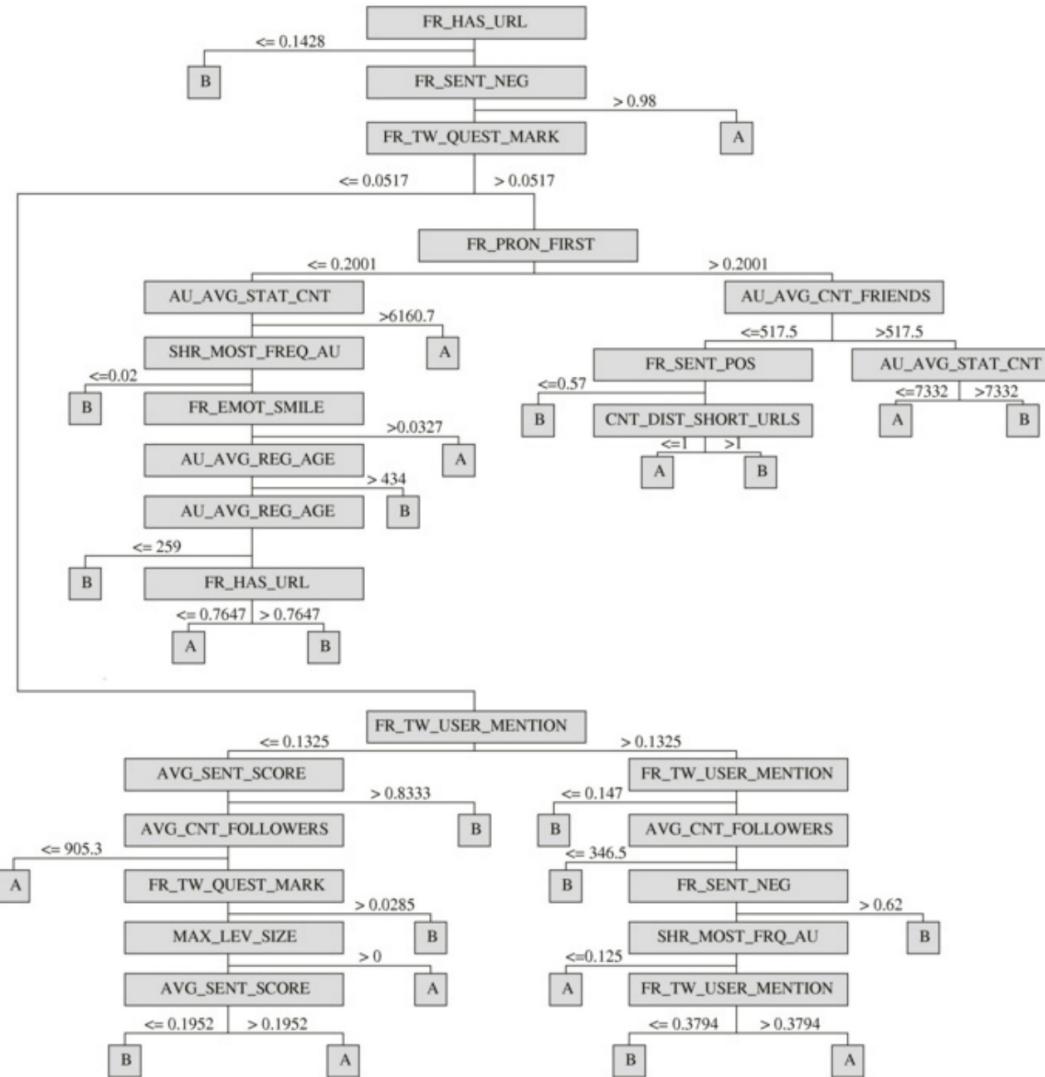


Interpretability

Being right for the right reasons



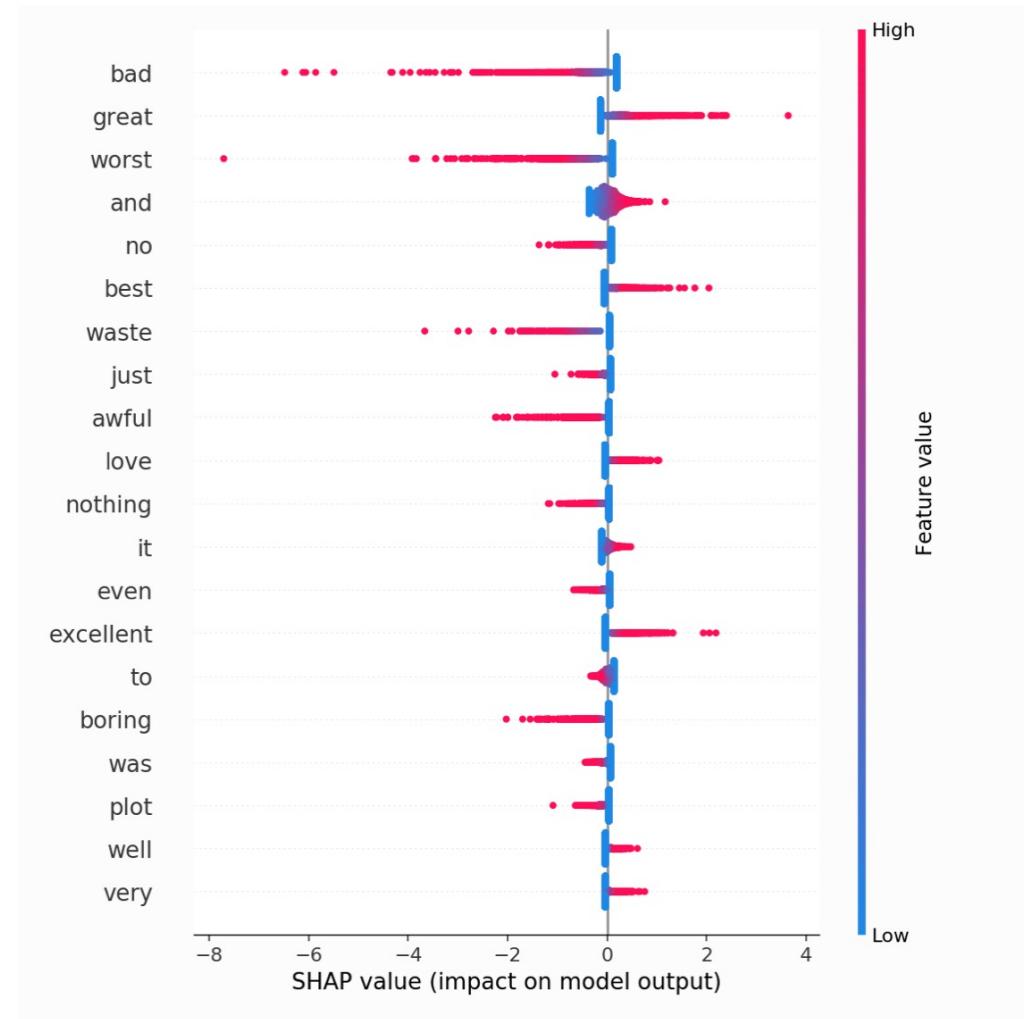
Interpretability: Model-dependent



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

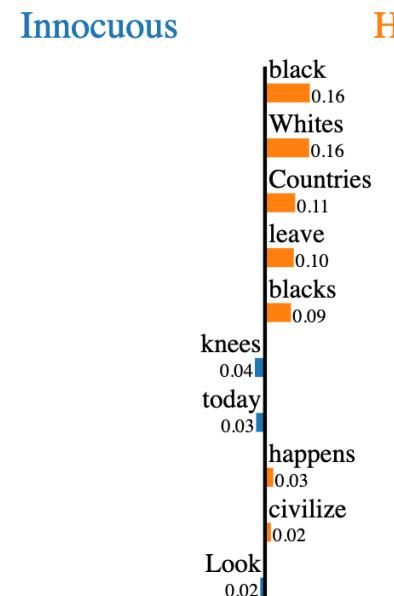
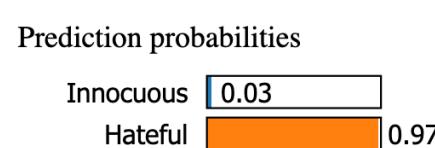
Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

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.



Text with highlighted words
Look what happens when **Whites** leave **black** **Countries** alone to do what they do naturally The **blacks** in White **Countries** today should be on their **knees** thanking **Whites** for trying to **civilize** them

Conclusion

- The Interpretation and Evaluation step is as important as the other steps in the text mining pipeline.
 - Applications
 - Responsible text mining
 - Interpretable / Explainable models

Practical (optional)
Focussing on application and
interpretation.

Questions?