FRAMING BIAS DETECTION

AI3011
MACHINE LEARNING
& PATTERN
RECOGNITION

PRESENTED BY :
GROUP 31

KANAV NANDA
SARTAJDEEP SINGH
SHLOK GUPTA

FRAMING BIAS DETECTION

PROBLEM STATEMENT

PROBLEM STATEMENT

WHAT IS THE PROBLEM?

THE INVISIBLE BIAS IN NEWS

News media doesn't have to lie to mislead. Most people consume news from mainstream sources where stories, though factually accurate, can still be biased through framing: the tone, perspective, or facts chosen to present an issue in a particular way.

This is framing bias and unlike fake news, it is virtually invisible to the casual reader.
Example: Same event, two headlines:

- "Demonstrators march for economic justice"
- "Angry mob disrupts city center"

Both can be factually defensible. Both paint entirely different pictures.

LINK : CHROME-EXTENSION://EFAIDNBMNNNIBPCAJPCGLCLEFINDMKAJ/HTTPS://ARXIV.ORG/PDF/2502.06009

PROBLEM STATEMENT

WHY CURRENT TOOLS FAIL?

THE BINARY TRAP

Most existing systems classify text as simply Biased or Not Biased, a binary output that:

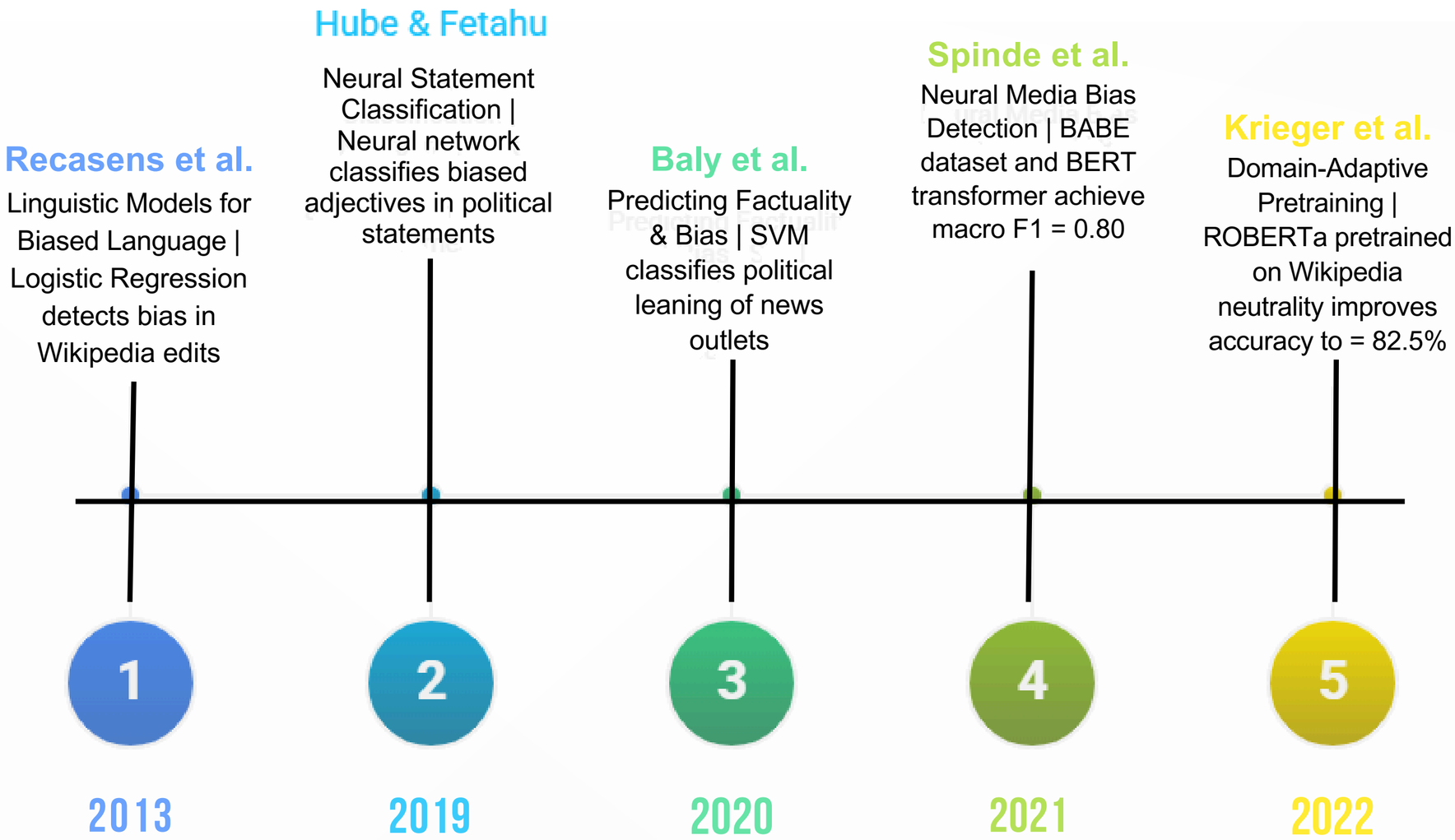
- Fails to communicate how much bias is present
- Cannot differentiate between mildly slanted wording and overtly inflammatory framing
- Gives no actionable insight to journalists, editors, or readers

The Core Gap: Bias exists on a spectrum, not a switch. A tool that cannot measure intensity cannot meaningfully help.

LITERATURE SURVEY

LITERATURE SURVEY

Evolution of Media Bias Detection Research (2013-2022)



RECASENS ET AL. (2013)

Linguistic Models for Analyzing and Detecting Biased Language

Marta Recasens
Stanford University
recasens@google.com

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

1. WHAT OTHER RESEARCHERS DID TO SOLVE THE PROBLEM?

Recasens et al. studied biased wording in Wikipedia articles and edits. They focused on identifying phrases that introduce subjective or non-neutral framing in text. The researchers analyzed linguistic patterns such as emotionally loaded words, hedges, and subjective expressions to distinguish biased language from neutral writing.

2. DEVELOPED SOLUTION FROM THE STUDY

The paper proposed a Logistic Regression–based linguistic model using lexical and syntactic features to detect biased phrasing. Features included:

- Subjective words
- Opinionated adjectives
- Hedges and intensifiers
- Part-of-speech patterns

Their approach demonstrated that linguistic cues can effectively identify framing bias in textual content.

3. HOW OUR PROJECT EXTENDS OR IMPROVES THIS WORK

- Uses a regression-based bias score instead of binary detection
- Extracts richer linguistic features:
- sentiment scores
- biased-word ratios
- adjective usage
- punctuation and capitalization patterns
- Improves overall prediction accuracy for headline framing bias.

https://web.stanford.edu/~jurafsky/pubs/neutrality.pdf

BALY ET AL. (2018\2019)

Predicting Factuality of Reporting and Bias of News Media Sources

Ramy Baly[1], **Georgi Karadzhov**[3], **Dimitar Alexandrov**[3], **James Glass**[1], **Preslav Nakov**[2]
[1]MIT Computer Science and Artificial Intelligence Laboratory, MA, USA
[2]Qatar Computing Research Institute, HBKU, Qatar;
[3]Sofia University, Bulgaria
{baly, glass}@mit.edu, pnakov@qf.org.qa
{georgi.m.karadjov, Dimityr.Alexandrov}@gmail.com

1. WHAT OTHER RESEARCHERS DID TO SOLVE THE PROBLEM?

Baly et al. investigated media bias at the news outlet level instead of individual headlines. They analyzed political leaning and factuality using information from: article content, Wikipedia pages, Twitter profiles, web traffic signals
The goal was to classify news organizations into political orientations such as left, center, or right.

2. DEVELOPED SOLUTION FROM THE STUDY

The study used Support Vector Machine (SVM) classifiers with textual and metadata features to predict:

- political bias
- factuality of reporting

The research showed that combining linguistic information with external metadata significantly improves media bias prediction.

3. HOW OUR PROJECT EXTENDS OR IMPROVES THIS WORK

- Our project improves upon this by:
- Detecting framing bias directly at the headline level, enabling fine-grained analysis.
- Introducing a continuous regression score rather than only categorical political labels.
- Using advanced linguistic and semantic features such as:
- biased-word ratios
- sentiment polarity
- entity detection
- adjective density
- hedge and intensifier counts
- Improving model performance through richer feature engineering and combined feature representations.

https://sls.csail.mit.edu/publications/2018/RamyBaly_EMNLP18.pdf

SPINDE ET AL. (2021)

Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts

Timo Spinde[1], Manuel Plank[2], Jan-David Krieger[2], Terry Ruas[1], Bela Gipp[1], Akiko Aizawa[3]

[1]University of Wuppertal, [2] University of Konstanz, [3] NII Tokyo
{firstname.lastname}@{uni-wuppertal.de,@uni-konstanz.de}
aizawa@nii.ac.jp

1. WHAT OTHER RESEARCHERS DID TO SOLVE THE PROBLEM?

Spinde et al. created the BABE (Bias Annotations By Experts) dataset to study media bias in news articles. They focused on detecting biased framing and wording using annotated political news content.

2. DEVELOPED SOLUTION FROM THE STUDY

The paper applied BERT-based transformer models for neural bias detection. Their transformer approach achieved strong performance with approximately:

- Macro F1-score ≈ 0.80

The study demonstrated that contextual language models can effectively capture subtle framing and ideological cues in political text.

3. HOW OUR PROJECT EXTENDS OR IMPROVES THIS WORK

- Our project builds upon transformer-based bias detection by:
- Combining textual embeddings with handcrafted linguistic features.
- Adding regression-based scoring to measure degree of framing bias.
- Engineering additional explainable features such as:
- punctuation counts
- capitalization usage
- adjective ratios
- sentiment and subjectivity measures
- named entity features
- These additions improve interpretability and help increase classification accuracy while providing more nuanced bias estimation.

https://media-bias-research.org/wp-content/uploads/2022/01/Neural_Media_Bias_Detection_Using_Distant_Supervision_With_BABE___Bias_Annotations_By_Experts_MBG.pdf

KRIEGER ET AL.

A Domain-adaptive Pre-training Approach for Language Bias Detection in News

Jan-David Krieger*
jan-david.krieger@uni-konstanz.de
University of Konstanz
Konstanz, Germany

Timo Spinde*
timo.spinde@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Terry Ruas
ruas@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Juhi Kulshrestha
juhi.kulshrestha@uni-konstanz.de
University of Konstanz
Konstanz, Germany

Bela Gipp
gipp@cs.uni-goettingen.de
University Göttingen
Göttingen, Germany

1. WHAT OTHER RESEARCHERS DID TO SOLVE THE PROBLEM?

Krieger et al. explored whether pretrained language models could better detect biased language if adapted to a neutrality-focused domain such as Wikipedia. They investigated how domain-specific pretraining influences bias classification performance.

2. DEVELOPED SOLUTION FROM THE STUDY

The researchers used Domain-Adaptive Pretraining (DAPT) on RoBERTa models using large-scale neutral Wikipedia text before fine-tuning on bias detection tasks. Their adapted transformer model improved performance and achieved:

- Accuracy ≈ 82.5%

The study showed that domain-specific pretraining improves generalization in media bias detection.

3. HOW OUR PROJECT EXTENDS OR IMPROVES THIS WORK

- Our project extends this research by:
- Combining deep textual representations with engineered linguistic and structural headline features.
- Using regression-based scoring to quantify framing intensity instead of only classification.
- Incorporating explainable NLP indicators such as:
- sentiment polarity
- subjectivity
- biased-word frequency
- POS-tag statistics
- named entity patterns
- This hybrid approach improves interpretability, captures subtle framing patterns, and contributes to higher predictive performance on news headline bias detection.

https://arxiv.org/pdf/2205.10773

Research Paper	Limitations
Recasens et al. (2013) *Linguistic Models for Detecting Biased Language*	• Focuses mainly on lexical and syntactic bias cues • Uses binary classification instead of bias intensity scoring • Limited semantic understanding of contextual framing • Does not combine embeddings with engineered linguistic features
Baly et al. (2019) *Predicting Factuality and Political Bias of News Media*	• Detects bias at outlet level rather than headline level • Relies heavily on metadata and external sources • Limited fine-grained framing analysis of individual headlines • Does not quantify degree of framing bias
Spinde et al. (2021) *Neural Media Bias Detection using BABE Dataset*	• Transformer models require large annotated datasets • Limited interpretability of deep learning predictions • Focuses mainly on classification rather than continuous scoring • Handcrafted linguistic indicators are not deeply explored
Krieger et al. (2022) *Domain-Adaptive Pretraining for Media Bias Detection*	• Primarily improves transformer pretraining strategy • Limited explainability of detected bias patterns • Focuses mainly on textual embeddings • Does not integrate detailed linguistic and sentiment-based features

DATASET

ABOUT THE DATASET

1. The **BABE** dataset contains **3,700 annotated news sentences**, making it significantly larger than many previous media bias datasets. Data was collected from **14 US news outlets** covering **12 controversial political and social topics from 2017–2020**. This improves both the topical diversity and political representativeness of the dataset.
2. The dataset includes **sentence-level bias labels, word-level bias annotations, and factual/opinion/mixed classifications**. Annotation was performed by multiple experts, with **8 annotators for 1,700 sentences and 5 annotators for the full dataset**. This multi-layer annotation process increases the dataset's credibility and usefulness for bias detection tasks.
3. The dataset achieved a **Krippendorff's Alpha of approximately 0.40**, showing relatively strong **inter-annotator agreement for a subjective task**. Previous media bias datasets often reported agreement scores between **0.00–0.21 only**. **The best benchmark model** trained on BABE **achieved a Macro F1-score of 0.804**, indicating strong label consistency and learnability.

ABOUT THE DATASET

RAW DATASET ANALYSIS

We started with a journalism-focused bias dataset containing the following key columns:

- text
- news_link
- outlet
- topic
- type
- label_bias
- label_opinion
- biased_words

BABE DATASET

text	news_link	outlet	topic	type	label_bias	label_opinion	biased_words
"Orange Is the New Black" star Yael Sto	https://www.foxnews.com/entertainment/australian-actress-yael-sto	Fox News	environment	right	Non-biased	Entirely factual	[]
"We have one beautiful law," Trump rec	https://www.alternet.org/2020/06/law-and-order-is-a-debased-conce	Alternet	gun control	left	Biased	Somewhat factual but also opinionated	['bizarre', 'characteristically']
...immigrants as criminals and eugenics,	https://www.nbcnews.com/news/latino/after-stephen-miller-s-white-	MSNBC	white-nationalism	left	Biased	Expresses writer's opinion	['criminals', 'fringe', 'extreme']
...we sounded the alarm in the early mo	https://www.alternet.org/2019/07/fox-news-has-gone-so-deep-into-v	Alternet	white-nationalism	left	Biased	Somewhat factual but also opinionated	[]
[Black Lives Matter] is essentially a non	http://feedproxy.google.com/~r/breitbart/~3/-vHhfcwC74U/	Breitbart	marriage-equality		Biased	Expresses writer's opinion	['cult']
[Democrats employ] their full arsenal to	https://thefederalist.com/2020/09/22/democrats-ramp-up-their-war-	Federalist	vaccine		Biased	Expresses writer's opinion	['apparent', 'crime']
[Newsoms's] obsession with masks has	https://thefederalist.com/2020/12/09/gavin-newsoms-covid-tyranny-i	Federalist	vaccine		Biased	Expresses writer's opinion	['obsession']
[Newsoms's] onslaught of propaganda i	https://thefederalist.com/2020/12/09/gavin-newsoms-covid-tyranny-i	Federalist	vaccine		Biased	Expresses writer's opinion	['propaganda', 'vilifying', 'unimpeded']
[The police] now prefer to think of then	http://feedproxy.google.com/~r/breitbart/~3/2RH0NVHEl6Q/	Breitbart	marriage-equality		Biased	Expresses writer's opinion	['wounds']
'A new low': Washington Post media cri	https://www.alternet.org/2019/08/a-new-low-washington-post-medi	Alternet	white-nationalism	left	Biased	Expresses writer's opinion	['blows', 'up', 'absurd', 'lies', 'nationalism
'Gangster capitalist' Trump is running a '	https://www.alternet.org/2020/07/gangster-capitalist-trump-is-runnin	Alternet	black lives matter	left	Biased	Somewhat factual but also opinionated	['Gangster', 'mafia', 'state', 'capitalist']
'The most progressive president since FI	https://feeds.feedblitz.com/~/629983109/0/alternet_all~%e2%80%98	Alternet	universal health care	left	Non-biased	Expresses writer's opinion	['pleasant']
'Woke' is not a dirty word but a moral d	https://www.breitbart.com/europe/2020/02/25/going-woke-is-a-mor	Breitbart	marriage-equality		Non-biased	Entirely factual	[]
"Harry Potter" author J.K. Rowling said	https://www.reuters.com/article/us-health-coronavirus-jkrowling/har	Reuters	blm		Non-biased	Entirely factual	[]
"He won because the Election was Rigg	https://www.reuters.com/article/us-usa-election/trump-concedes-no	Reuters	vaccine		Non-biased	Entirely factual	[]
"I would shut it down; I would listen to '	https://thefederalist.com/2020/08/22/biden-i-will-shut-down-nation-i	Federalist	vaccine		Non-biased	Entirely factual	[]
"Mike Pence and the task force have do	https://www.reuters.com/article/us-health-coronavirus-usa-task-forc	Reuters	vaccine		Non-biased	Entirely factual	[]
"She will be inheriting a government in c	https://www.breitbart.com/politics/2020/01/23/elizabeth-warren-vov	Breitbart	marriage-equality		Biased	Somewhat factual but also opinionated	[]
"The goal is to send a message of peace	http://feedproxy.google.com/~r/breitbart/~3/KernUHEqyEk/	Breitbart	black lives matter	right	Non-biased	Somewhat factual but also opinionated	['claimed']
"We can't keep our country closed for t	https://www.reuters.com/article/us-health-coronavirus-usa-task-forc	Reuters	vaccine		Non-biased	Entirely factual	[]
"You know, there's over 100 million pec	http://feedproxy.google.com/~r/breitbart/~3/hmjqwWAvPXQ/	Breitbart	universal health care	right	Biased	Somewhat factual but also opinionated	['claiming']
23 people were arrested for offences in	https://www.breitbart.com/europe/2020/06/01/23-arrested-london-b	Breitbart	black lives matter	right	Non-biased	Entirely factual	[]
A 10-hour hearing broadcast on the cou	https://www.foxnews.com/sports/olympic-swimming-sun-yang-banne	Fox News	sport	right	Non-biased	Somewhat factual but also opinionated	[]
A 16-time NBA All Star who is considere	https://www.reuters.com/article/us-usa-election-james/nba-star-lebr	Reuters	black lives matter	center	Non-biased	Entirely factual	[]
A 24-year-old woman in Uganda, unable	https://www.foxnews.com/world/coronavirus-persecution-islam-chris	Fox News	islam		Non-biased	Entirely factual	[]
A 60-foot chunk of the Ocean Cleanup c	https://eu.usatoday.com/story/news/2019/01/02/ocean-cleanup-dev	USA Today	environment	center	Non-biased	Entirely factual	['fanfare']

***3675 ROWS**

1810 BIASED, 1863 NON BIASED

DATASET: https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE

CLASSIFICATION MODEL

DATA PREPROCESSING

Dataset Normalization

- Normalized the complete dataset to maintain uniform feature scaling.
- Prevented high-magnitude features from dominating the training process.

Label Encoding

- Converted categorical bias labels into numerical values:
  - Biased → 1
  - Unbiased → 0

Removing Ambiguous Samples

- Removed rows where:
  - label_opinion = "No Agreement"
- Eliminated conflicting expert annotations and reduced label noise.

Dataset Cleaning

- Removed noisy, duplicate, and incomplete samples.

DATA PREPROCESSING

WE EXTRACTED MULTIPLE LINGUISTIC, LEXICAL, AND SEMANTIC FEATURES:

TEXT STATISTICS

- char_len
- word_count
- avg_word_len
- punct_count
- caps_word_count

LINGUISTIC STRUCTURE FEATURES

- adj_count
- adj_ratio
- verb_count
- noun_count

SEMANTIC/NER FEATURES

- entity_count
- has_person
- has_org
- has_place
- person_count

TONE & FRAMING FEATURES

- intensifier_count
- hedge_count
- negative_word_count
- is_question
- is_exclamation

BIAS-ORIENTED FEATURES

- biased_word_count
- biased_word_ratio
- biased_word_present

SENTIMENT FEATURES

- vader_compound
- vader_pos
- vader_neg
- vader_neu
- subjectivity
- polarity

DATA PREPROCESSING

GLOSSARY

Feature	Meaning
char_len	Character Length
word_count	Total Word Count
avg_word_len	Average Word Length
punct_count	Punctuation Count
caps_word_count	Capitalized Word Count
biased_word_count	Count of Biased Words
biased_word_ratio	Ratio of Biased Words
biased_word_present	Presence of Biased Words
vader_compound	VADER Compound Score
vader_pos	VADER Positive Score
vader_neg	VADER Negative Score
vader_neu	VADER Neutral Score
subjectivity	Subjectivity Score
polarity	Sentiment Polarity Score

Feature	Meaning
adj_count	Adjective Count
adj_ratio	Adjective Ratio
verb_count	Verb Count
noun_count	Noun Count
entity_count	Named Entity Count
has_person	Presence of Person Entity
has_org	Presence of Organization Entity
has_place	Presence of Place Entity
person_count	Person Entity Count
intensifier_count	Intensifier Word Count
hedge_count	Hedge Word Count
negative_word_count	Negative Word Count
is_question	Question Indicator
is_exclamation	Exclamation Indicator

FINAL DATASET

I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
target_sc	char_len	word_cou	avg_word	punct_cou	caps_wor	biased_w	biased_w	biased_w	vader_cor	vader_pos	vader_neg	vader_neu	subjectiv	polarity	intensifie	hedge_co	negative_	is_questi	is_exclam	entity_co	has_pers	has_org	has_place	person_c	adj_count	adj_ratio	verb_cou	noun_count
1	0.232172	0.255102	0.288462	0.375	0.166667	0	0	0	0.30645	0	0.172477	0.848387	0.39596	0.512907	0	0	0	0	0	0.333333	1	0	1	0.2	0.153846	0.102564	0.214286	0.192308
8.5	0.217247	0.183673	0.582237	0.125	0	0.125	0.210526	1	0.890792	0.483819	0.161468	0.375806	0.7125	0.847222	0	0	0	0	0	0.166667	0	1	0	0	0.230769	0.210526	0.214286	0.153846
10	0.271973	0.22449	0.619565	0.625	0	0.1875	0.26087	1	0.419094	0.236246	0.244037	0.548387	0.25	0.538194	0	0	0	0	0	0	0	0	0	0	0.307692	0.231884	0.142857	0.307692
6.5	0.525705	0.5	0.45375	0.5	0	0	0	0	0.240412	0.072816	0.192661	0.758065	0.34	0.666667	0	0	0.5	0	0	0.333333	1	1	0	0.2	0.461538	0.16	0.357143	0.538462
8.95	0.25539	0.204082	0.660714	0.125	0	0.0625	0.095238	1	0.697959	0.257282	0	0.743548	0.245833	0.532407	0	0	0	0	0	0.166667	0	0	0	0	0.384615	0.31746	0.142857	0.076923
10	0.205638	0.183673	0.513158	0.125	0	0.125	0.210526	1	0.488977	0.229773	0.26789	0.533871	0.55	0.722222	0	0	0	0	0	0.25	1	0	0	0.2	0.153846	0.140351	0.214286	0.192308
9.33	0.174129	0.142857	0.6125	0.125	0	0.0625	0.133333	1	0.295427	0.173139	0.462385	0.420968	0	0.555556	0	0	0	0	0	0.083333	0	0	0	0	0.076923	0.088889	0.071429	0.230769
9.54	0.421227	0.387755	0.490385	0.125	0	0.1875	0.153846	1	0.266971	0.171521	0.390826	0.485484	0.194444	0.484568	0	0	0	0	0	0.083333	0	0	0	0	0.076923	0.034188	0.642857	0.423077
8.43	0.482587	0.459184	0.452446	0.125	0	0.0625	0.043478	1	0.744001	0.312298	0.288073	0.435484	0.3875	0.631944	0	0	0	0	0	0.166667	0	1	0	0	0.076923	0.028986	0.714286	0.346154
10	0.165837	0.153061	0.46875	0	0.166667	0.375	0.75	1	0.194883	0	0.475229	0.582258	0.438636	0.505051	0	0	0	0	0	0.166667	1	1	0	0.2	0.230769	0.25	0.071429	0.192308
8.5	0.202322	0.183673	0.493421	0	0	0.25	0.421053	1	0.257793	0	0.269725	0.762903	0.6	0.333333	0	0	0	0	0	0.083333	1	0	0	0.2	0.153846	0.140351	0.214286	0.269231
5.54	0.15257	0.122449	0.634615	0	0.166667	0.0625	0.153846	1	0.840443	0.532362	0	0.469355	0.733333	0.898148	0	0	0	0	0	0.166667	1	0	0	0.4	0.153846	0.205128	0.071429	0.153846
1	0.384743	0.377551	0.414474	0.125	0.166667	0	0	0	0.80281	0.190939	0	0.809677	0.289583	0.611111	0	0	0	0	0	0.166667	0	1	0	0	0.615385	0.280702	0.285714	0.269231
1	0.208955	0.214286	0.357955	0.375	0.333333	0	0	0	0.216417	0	0.343119	0.698387	0	0.555556	0	0	0	0	0	0.416667	1	1	0	0.4	0	0	0.214286	0.192308
1	0.218905	0.22449	0.358696	0.125	0	0	0	0	0.630384	0.218447	0.181651	0.622581	0	0.555556	0	0	0	0	0	0.166667	1	0	0	0.4	0	0	0.357143	0.192308
1	0.257048	0.255102	0.396635	0.25	0.333333	0	0	0	0.502205	0	0	1	0.371717	0.550224	0	0	0	0	0	0.333333	1	1	1	0.4	0.153846	0.102564	0.214286	0.153846
1	0.177446	0.214286	0.196023	0.125	0	0	0	0	0.8226	0.263754	0	0.737097	0.75	1	0	0	0	0	0	0.25	1	0	1	0.4	0.076923	0.060606	0.142857	0.230769
6.5	0.230514	0.214286	0.46875	0.125	0	0	0	0	0.276405	0	0.299083	0.737097	0.6	0.333333	0	0	0	0	0	0.083333	0	1	0	0	0.076923	0.060606	0.214286	0.269231
2.97	0.401327	0.428571	0.322674	0.25	0	0.0625	0.046512	1	0.121257	0.074434	0.278899	0.680645	0.575	0.659722	0	0	0	0	0	0.166667	1	0	0	0.2	0.307692	0.124031	0.428571	0.5
1	0.205638	0.244898	0.21	0.125	0	0	0	0	0.502205	0	0	1	0.12963	0.50823	0	0	0	0	0	0.166667	1	0	0	0.2	0.076923	0.053333	0.357143	0.192308
7.27	0.281924	0.255102	0.504808	0.125	0	0.0625	0.076923	1	0.677143	0.142395	0	0.858065	0.625	0.833333	0	0	0	0	0	0.083333	0	0	0	0	0.076923	0.051282	0.5	0.192308
1	0.237148	0.214286	0.502841	0.125	0	0	0	0	0.078138	0.063107	0.577982	0.427419	0.433333	0.462963	0	0	0	0	0	0.416667	0	1	1	0	0.076923	0.060606	0.214286	0.230769
2.5	0.404643	0.387755	0.442308	0.125	0	0	0	0	0.653968	0.153722	0.165138	0.7	0.333333	0.740741	0	0	0	0	0	0.5	0	1	1	0	0.307692	0.136752	0.214286	0.423077
1	0.348259	0.367347	0.334459	0.125	0.166667	0	0	0	0.676528	0.161812	0.122936	0.730645	1	0.694444	0	0	0	0	0	0.833333	1	1	1	0.2	0.153846	0.072072	0.214286	0.307692
1	0.325041	0.285714	0.543103	0.125	0	0	0	0	0.502205	0	0	1	0.4375	0.381944	0	0	0	0	0	0.333333	0	0	1	0	0.153846	0.091954	0.285714	0.307692
1.49	0.383085	0.408163	0.324695	0.125	0	0.0625	0.04878	1	0.630384	0.144013	0.122936	0.748387	0.395	0.622222	0	0	0	0	0	0.333333	0	1	0	0	0.153846	0.065041	0.214286	0.384615
1	0.39801	0.367347	0.486486	0.125	0	0	0	0	0.834752	0.213592	0	0.787097	0.621429	0.584656	0	0.5	0	0	0	0.25	0	1	0	0	0.307692	0.144144	0.357143	0.384615
10	0.101161	0.091837	0.46875	0.125	0	0.125	0.4	1	0.502205	0	0	1	0	0.555556	0.5	0	0	0	0	0.083333	0	1	0	0	0.153846	0.266667	0	0.153846
10	0.18408	0.183673	0.384868	0.125	0	0.125	0.210526	1	0.158429	0	0.447706	0.606452	0	0.555556	0	0	0	0	0	0.083333	0	1	0	0	0.461538	0.421053	0.071429	0.192308
10	0.225539	0.153061	0.890625	0.125	0	0.25	0.5	1	0.697959	0.239482	0	0.76129	0	0.555556	0	0	0	0	0	0.166667	0	1	0	0	0.923077	1	0.142857	0.230769
8	0.150912	0.142857	0.4375	0.125	0	0	0	0	0.208983	0	0.383486	0.662903	0.84375	0.46875	0.5	0	0	0	0	0.166667	0	1	1	0	0.076923	0.088889	0.142857	0.153846
1	0.311774	0.295918	0.45	0	0	0	0	0	0.169658	0	0.236697	0.791935	0.4	0.444444	0	0	0	0	0	0.25	0	1	0	0	0.153846	0.088889	0.214286	0.269231

CLASSIFICATION

We implemented four machine learning models to classify news headlines as biased or non-biased:

- **Logistic Regression (LR):** Linear classification model with GridSearchCV hyperparameter tuning using 5-fold cross-validation.
- **XGBoost Classifier (XGB):** Gradient boosting model with 300 decision trees for learning complex non-linear relationships.
- **Random Forest Classifier (RF):** Ensemble tree-based model using 300 trees to improve robustness and reduce overfitting.
- **Support Vector Machine (SVM)**: Linear SVM with probability calibration for confidence-based classification.

Feature Sets Evaluated

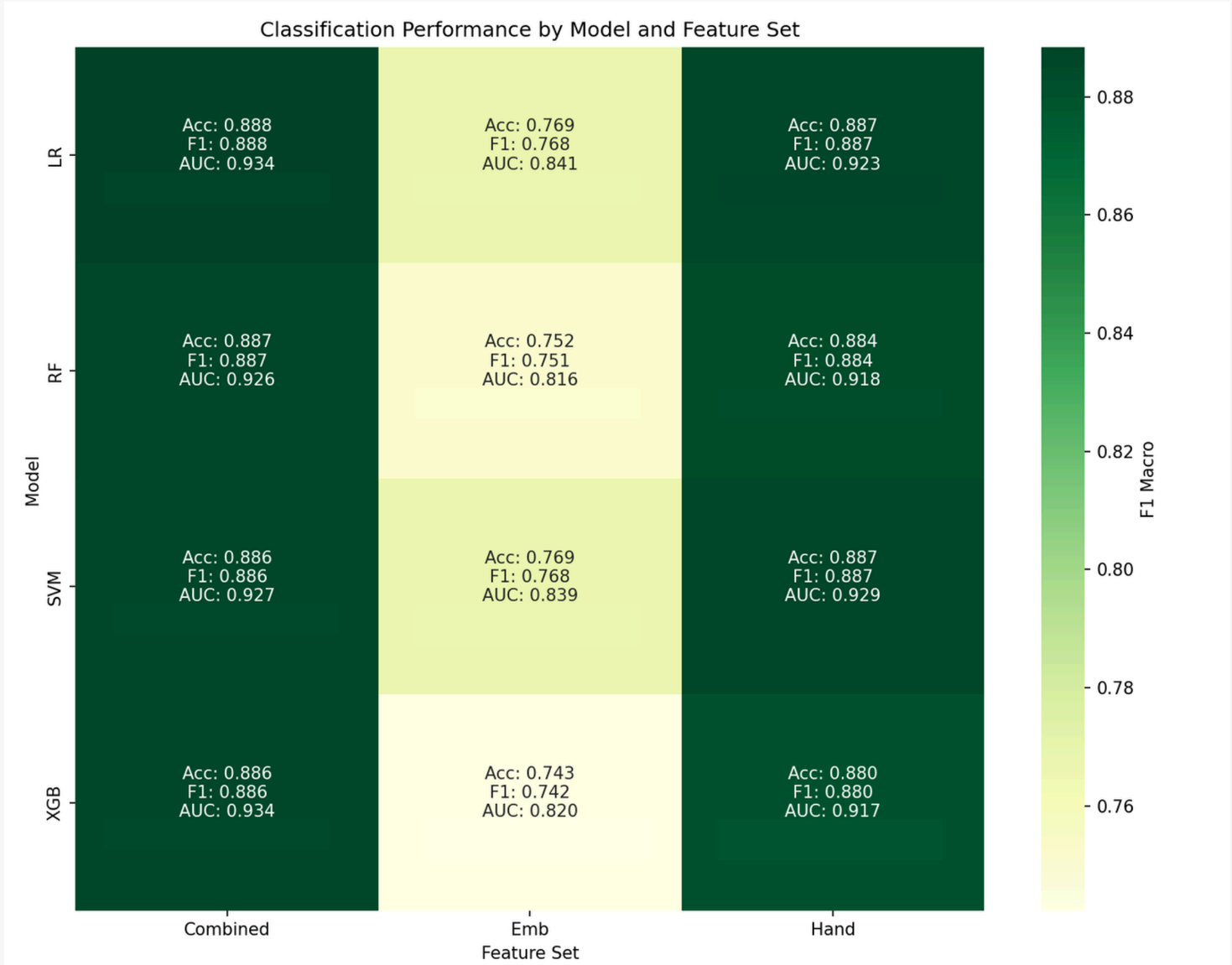
- Emb: Sentence-transformer embeddings
- Hand: Handcrafted linguistic and sentiment-based features
- Combined: Embeddings + handcrafted NLP features

Evaluation Metrics

Models were evaluated using:

- Accuracy
- F1 Macro Score
- ROC-AUC

CLASSIFICATION



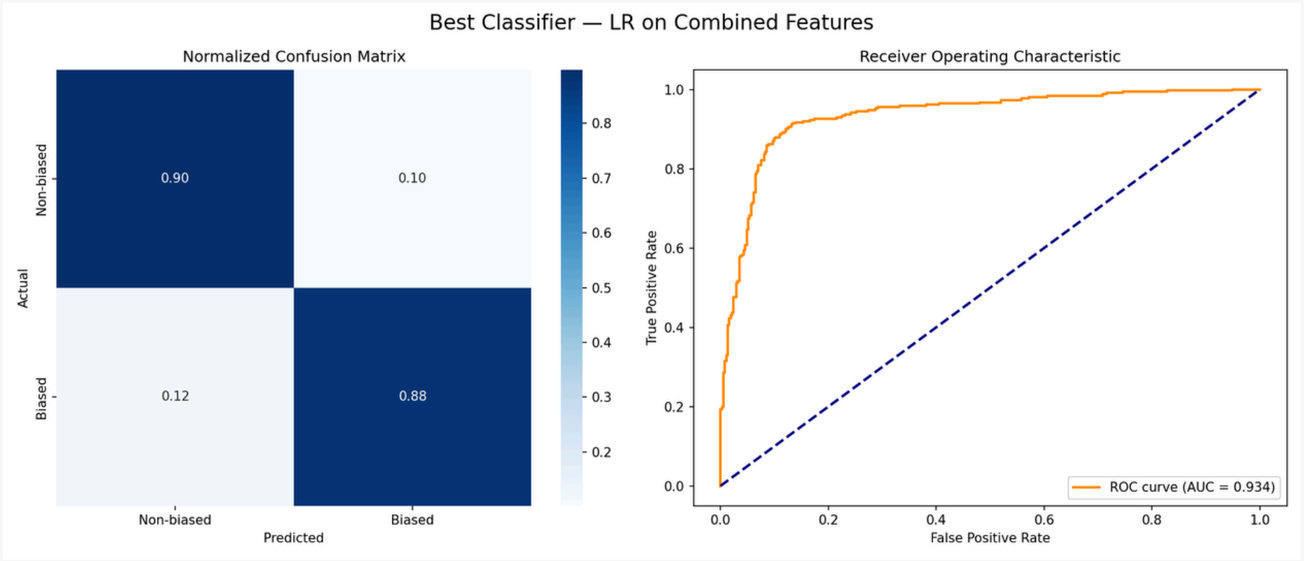
CLASSIFICATION

- After extensive feature engineering, we wanted to validate whether the extracted features were genuinely informative for bias detection or simply adding noise to the dataset.
- To verify whether the engineered features were relevant and informative, we trained a Logistic Regression classifier on the fully processed dataset.

Performance Analysis

The normalized confusion matrix showed:

- 90% correct classification for Non-Biased samples
- 88% correct classification for Biased samples
- **Accuracy : 0.89**
- **Precision : 0.898**
- **AUC Score = 0.934**
- **F1 SCORE = 0.888**



Key Findings

- The model achieved strong classification performance.
- ROC analysis showed good class separability.
- Validated that the engineered features.

The strongest benchmark in the literature was BERT with distant supervision, achieving a Macro F1-score of 0.804 while our models F1 score being 0.888 is better.

FURTHERMORE

The dataset only provided binary/categorical bias annotations, which was insufficient for our initial goal of measuring bias severity on a continuous scale (1–10).

SO OUR NEW APPROACH INVOLVES INTRODUCING

REGRESSION SCORE BASED MODEL

This will transform the problem from:
Binary Classification to Continuous Regression

OUR GOLDEN FORMULA

$$S = 1.0 + (B \times 4.0) + (O \times 3.0) + \left(\frac{W}{L} \times 2.0\right)$$

1. **Base (1.0)** The absolute minimum score, representing a completely neutral and factual sentence.
2. **Expert Bias (B) | Weight: +4.0** A binary flag indicating if trained media experts explicitly labeled the text as biased. If experts agree it is biased, it mathematically guarantees a high severity score.
3. **Opinion/Subjectivity (O)| Max Weight: +3.0** A tiered multiplier based on the expert's classification of how the information is presented . "**Entirely factual**" = **+0.0** (No penalty, purely objective). "**Somewhat factual but also opinionated**" = **+1.5** (Partial penalty for mixing facts with subjective framing). "**Expresses writer's opinion**" = **+3.0** (Maximum penalty for entirely subjective claims masked as news).
4. **Lexical Density (W/L) | Weight: +2.0** The ratio of loaded/biased words to the total length of the sentence. A short sentence packed with slanted words gets a higher intensity penalty than a long paragraph with a single loaded word.

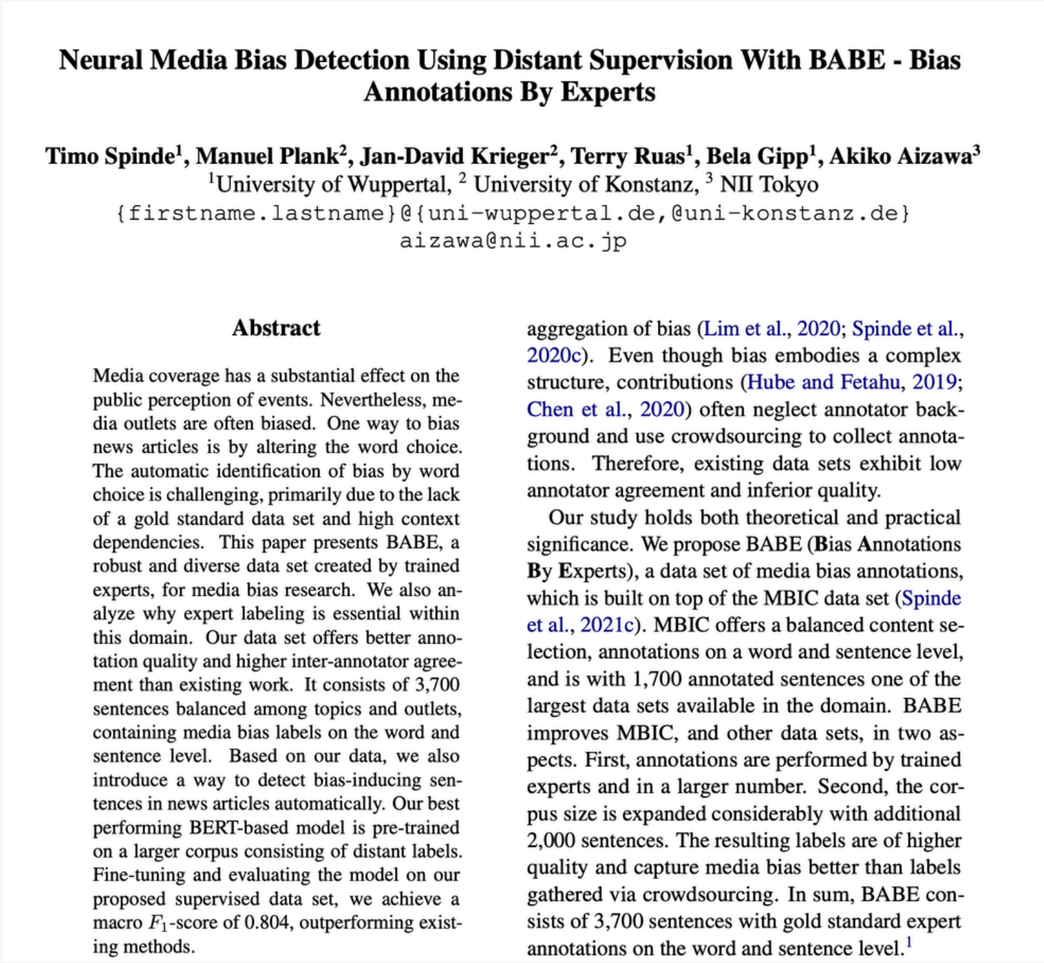
WHY THIS WORKS

Prior research in "Neural Media Bias Detection Using Distant Supervision With BABE" argues that media bias is highly contextual and difficult to capture using simple automated or crowdsourced labeling methods.
The paper specifically highlights that crowdsourced annotations often produce low annotator agreement expert annotations are significantly higher quality and more reliable

Trained experts better identify subtle framing and ideological manipulation
BABE therefore constructs a dataset based on gold-standard expert annotations for both sentence-level and word-level bias detection.

CONNECTION TO OUR FORMULA:

As prior literature treats expert- onfirmed bias as the most reliable indicator of actual media bias, our framework assigns the largest weight (+4.0) to expert-labeled bias.
This ensures that expert consensus strongly influences the bias severity score.



Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts

Timo Spinde[1], Manuel Plank[2], Jan-David Krieger[2], Terry Ruas[1], Bela Gipp[1], Akiko Aizawa[3]
[1]University of Wuppertal, [2] University of Konstanz, [3] NII Tokyo
{firstname.lastname}@{uni-wuppertal.de,@uni-konstanz.de}
aizawa@nii.ac.jp

Abstract

Media coverage has a substantial effect on the public perception of events. Nevertheless, media outlets are often biased. One way to bias news articles is by altering the word choice. The automatic identification of bias by word choice is challenging, primarily due to the lack of a gold standard data set and high context dependencies. This paper presents BABE, a robust and diverse data set created by trained experts, for media bias research. We also analyze why expert labeling is essential within this domain. Our data set offers better annotation quality and higher inter-annotator agreement than existing work. It consists of 3,700 sentences balanced among topics and outlets, containing media bias labels on the word and sentence level. Based on our data, we also introduce a way to detect bias-inducing sentences in news articles automatically. Our best performing BERT-based model is pre-trained on a larger corpus consisting of distant labels. Fine-tuning and evaluating the model on our proposed supervised data set, we achieve a macro F_1-score of 0.804, outperforming existing methods.

aggregation of bias (Lim et al., 2020; Spinde et al., 2020c). Even though bias embodies a complex structure, contributions (Hube and Fetahu, 2019; Chen et al., 2020) often neglect annotator background and use crowdsourcing to collect annotations. Therefore, existing data sets exhibit low annotator agreement and inferior quality.

Our study holds both theoretical and practical significance. We propose BABE (**B**ias **A**nnotations **B**y **E**xperts), a data set of media bias annotations, which is built on top of the MBIC data set (Spinde et al., 2021c). MBIC offers a balanced content selection, annotations on a word and sentence level, and is with 1,700 annotated sentences one of the largest data sets available in the domain. BABE improves MBIC, and other data sets, in two aspects. First, annotations are performed by trained experts and in a larger number. Second, the corpus size is expanded considerably with additional 2,000 sentences. The resulting labels are of higher quality and capture media bias better than labels gathered via crowdsourcing. In sum, BABE consists of 3,700 sentences with gold standard expert annotations on the word and sentence level.[1]

WHY THIS WORKS

Prior work in "Sentiment Analysis in the News" shows that news bias is closely tied to how information is presented rather than only what information is presented.
The paper distinguishes between: factual reporting, partially opinionated reporting, explicitly subjective or evaluative statements.

It argues that opinion mining in news differs from normal sentiment analysis because journalistic bias often appears through subtle subjective framing embedded within factual content.

CONNECTION TO OUR FORMULA:

This directly motivates our tiered subjectivity component:
factual statements receive no penalty,
partially opinionated statements receive a moderate increase,
explicitly opinion-based reporting receives the maximum increase.

Sentiment Analysis in the News

Alexandra Balahur[1], Ralf Steinberger[2], Mijail Kabadjov[2], Vanni Zavarella[2], Erik van der Goot[2], Matina Halkia[2], Bruno Pouliquen[2], Jenya Belyaeva[2]

[1] University of Alicante, Department of Software and Computing Systems
Ap. de Correos 99, E-03080 Alicante, Spain
[2] European Commission – Joint Research Centre
IPSC - GlobeSec - OPTIMA (OPensource Text Information Mining and Analysis)
T.P. 267, Via Fermi 2749
21027 Ispra (VA), Italy

abalahur@dlsi.ua.es,
{Ralf.Steinberger, Mijail.Kabadjov, Erik.van-der-Goot, Matina.Halkia,Bruno.Pouliquen}@jrc.ec.europa.eu,
{Vanni.Zavarella, Jenya.Belyaeva}@ext.jrc.ec.europa.eu

Abstract

Recent years have brought a significant growth in the volume of research in sentiment analysis, mostly on highly subjective text types (movie or product reviews). The main difference these texts have with news articles is that their target is clearly defined and unique across the text. Following different annotation efforts and the analysis of the issues encountered, we realised that news opinion mining is different from that of other text types. We identified three subtasks that need to be addressed: definition of the target; separation of the good and bad news content from the good and bad sentiment expressed on the target; and analysis of clearly marked opinion that is expressed explicitly, not needing interpretation or the use of world knowledge. Furthermore, we distinguish three different possible views on newspaper articles – author, reader and text, which have to be addressed differently at the time of analysing sentiment. Given these definitions, we present work on mining opinions about entities in English language news, in which (a) we test the relative suitability of various sentiment dictionaries and (b) we attempt to separate positive or negative opinion from good or bad news. In the experiments described here, we tested whether or not subject domain-defining vocabulary should be ignored. Results showed that this idea is more appropriate in the context of news opinion mining and that the approaches taking this into consideration produce a better performance.

WHY THIS WORKS

Prior research in "Linguistic Models for Analyzing and Detecting Biased Language" identifies several linguistic patterns strongly associated with biased reporting, including intensifiers, one-sided terms,assertives,hedges,factive verbs.

The paper further shows that framing bias constitutes a major portion of observed media bias, with "one-sided terms" being one of the most dominant subcategories.

This suggests that biased reporting frequently manifests through concentrated framing vocabulary and emotionally loaded wording.

CONNECTION TO OUR FORMULA:

Based on this finding, our framework incorporates lexical density as a measure of framing intensity:

$$LexicalDensity = \frac{\text{Biased Words}}{\text{Sentence Length}}$$

A higher concentration of loaded language increases the media bias score because prior literature identifies framing vocabulary as a major linguistic signal of bias.

Linguistic Models for Analyzing and Detecting Biased Language

Marta Recasens
Stanford University
recasens@google.com

Cristian Danescu-Niculescu-Mizil
Stanford University
Max Planck Institute SWS
cristiand@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Bias	Subtype	%
A. Epistemological bias		43
	- Factive verbs	3
	- Entailments	25
	- Assertives	11
	- Hedges	4
B. Framing bias		57
	- Intensifiers	19
	- One-sided terms	38

Table 2: Proportion of the different bias types.

METHODOLOGY

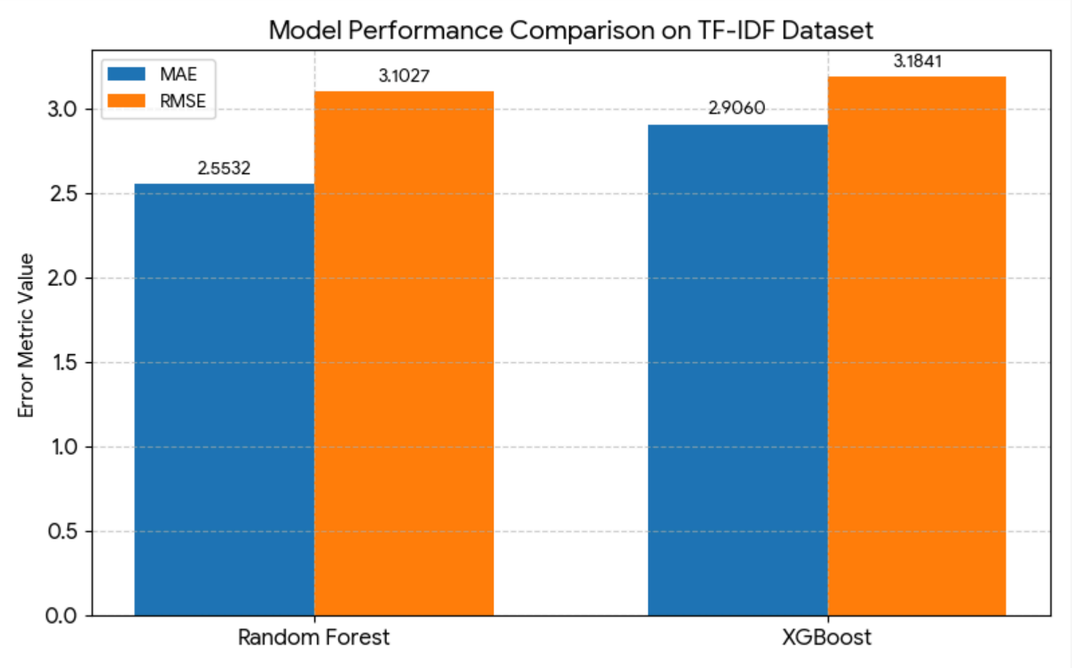
METHODOLOGY - 1

ML Approach: TF-IDF

We Applied TF-IDF on the complete textual dataset to convert news text into numerical vectors.

Results -

Model	MAE	RMSE
Random Forest Regressor	2.5532	3.1027
XGBoost Regressor	2.9060	3.1841



Key Observation

Although TF-IDF captured keyword importance, it:

- Failed to capture contextual semantics
- Ignored sentence structure
- Could not model nuanced framing bias effectively

Conclusion

Traditional bag-of-words representations were insufficient for subtle bias severity estimation

TF/IDF

original_te	target_sco	10	100	1000	10000	100000	11	12	13	14	15	150	16	1619	17	18	19	1964	1969	20	200	2000
A 24-year-	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A 60-foot c	1.49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A 70% maj	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Biden vic	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Biden vic	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Biden vic	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Biden vic	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A bill propc	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A bipartisa	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Black ma	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Black ma	1	0	0	0	0	0	0	0	0	0	0.218147	0	0	0	0	0	0	0	0	0.200251	0	0
A Black ma	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Black pol	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A bold stat	7.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Catholic	2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A city coun	5.49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A code tha	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A cop shoc	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A copy of th	5.67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A cursoryÂ	8.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A Democra	5.61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A diverse g	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A divided fe	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A donation	6.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A family-ru	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

METHODOLOGY - 2

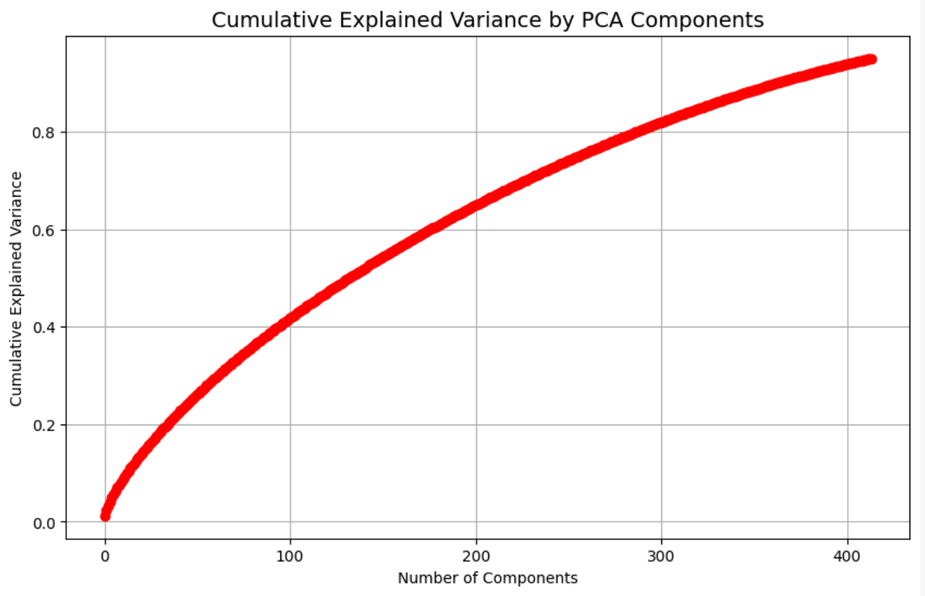
ML Approach: Dimensionality Reduction using PCA

Problem with Previous Pipeline

- Extremely high-dimensional sparse vectors
- Redundant/noisy features
- Increased computational complexity

Solution: Principal Component Analysis (PCA)

- Reduce dimensionality
- Remove redundant variance



PCA Statistics

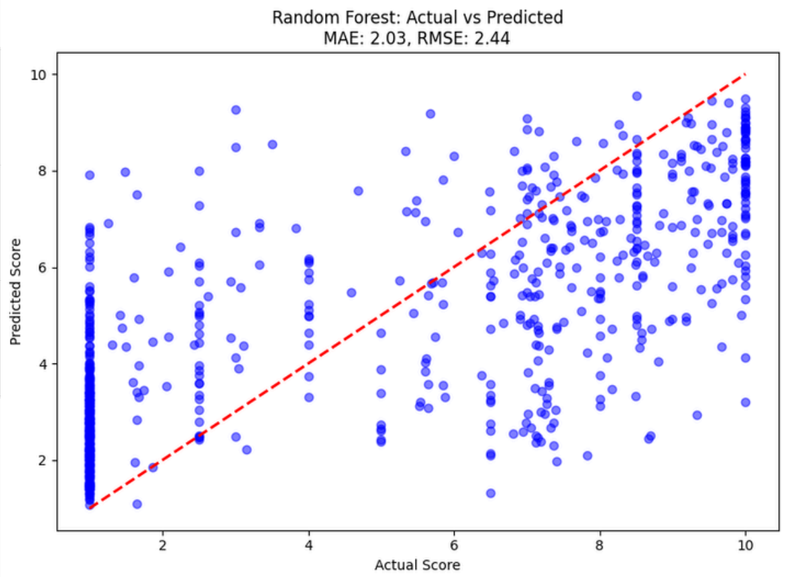
Metric	Value
Original Feature Count	528
Reduced Feature Count	414
Variance Retained	95%

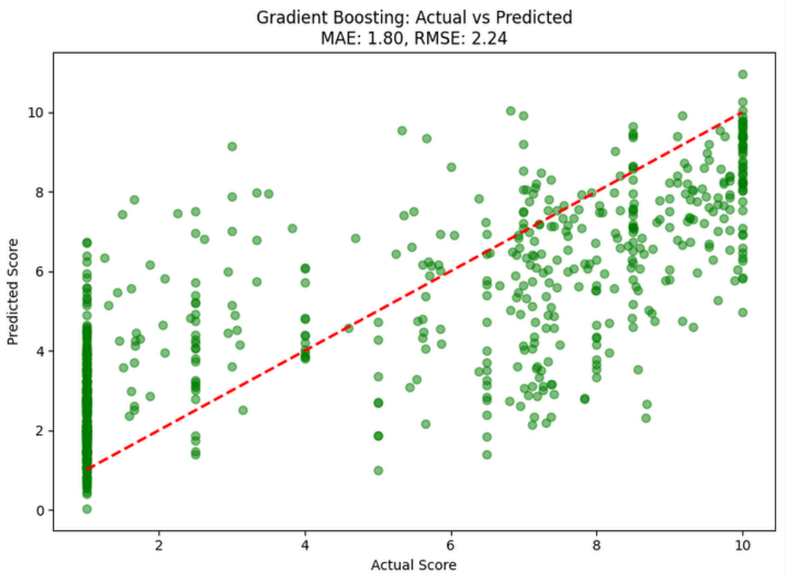
METHODOLOGY - 2

ML Approach: Dimensionality Reduction using PCA

Retrained Models on PCA Features

MODEL	MAE	RMSE
RANDOM FOREST	2.032	2.444
XGBoost	1.803	2.239





PCA significantly improved:

- Noise reduction
- Generalization capability
- Regression stability

Key Insight

Reducing irrelevant feature variance allowed XGBoost to focus on meaningful framing patterns.

METHODOLOGY - 3

ML Approach: Embedding-Based Approach

TF-IDF still lacked:

- Semantic understanding
- Context awareness
- Relationship between words

Proposed Improvement -

Text Embeddings

Headline embeddings were generated using the Sentence-Transformer model all-mpnet-base-v2, which converts each headline into dense semantic vector representations.

Feature Sets Evaluated

- Emb: Transformer embeddings only
- Hand: Handcrafted NLP features only
- Combined: Embeddings + engineered features

EMBEDDINGS

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
-0.001461	0.139129	0.014495	-0.02053	0.058245	0.007134	0.008255	0.020406	-0.019157	0.016874	0.019418	0.01096	0.007184	0.041866	-0.071805	0.013889	-0.013709	-0.016449	0.064225	-0.04121	-0.049114	0.026083	0.029681	0.041824	-0.055181	-0.031139	0.013531	0.036808
0.046131	0.062914	-0.010792	-0.017995	-0.055486	0.010924	-0.043664	0.049497	-0.020782	-0.029761	0.035114	-0.018805	0.007107	-0.046198	0.004991	-0.041564	0.042275	0.016923	0.046156	-0.043445	-0.016	0.007504	0.035034	-0.029603	0.021279	0.019904	0.027332	0.010122
-0.030679	0.005186	-0.019589	0.012654	-0.067262	0.001337	0.020581	-0.031335	0.051311	0.003299	-0.04251	-0.017268	0.001588	-0.044006	0.009615	-0.002377	0.011945	0.003716	0.003317	-0.003053	0.000261	0.02809	-0.013261	0.005219	-0.022187	-0.001286	-0.002815	-0.066832
-0.055316	0.071214	-0.000922	0.003523	-0.016205	0.043549	0.014357	-0.013679	-0.018102	-0.080499	-0.005137	0.013373	-0.012402	-0.013485	0.030431	0.044597	0.038351	0.043897	0.031765	-0.005789	-0.038613	0.019374	-0.010446	0.008421	-0.045238	-0.021913	0.042439	-0.013919
-0.00276	-0.006601	0.017762	0.088037	-0.064026	0.006387	0.012224	-0.040875	-0.024165	-0.004809	0.009719	0.002316	0.001767	-0.000861	0.027955	0.076099	-0.017981	0.025321	-0.047393	0.031743	-0.012857	0.041263	0.001334	-0.010449	0.026942	-0.001438	0.000894	0.013242
0.012574	0.023107	0.004754	-0.005064	-0.053686	0.01039	0.018576	-0.008456	0.009044	-0.072624	-0.000835	0.036368	-0.008997	-0.003893	0.029682	0.014715	0.014123	0.029004	-0.025788	-0.016456	-0.0198	-0.022041	0.007827	-0.026806	-0.005942	-0.041307	0.010036	0.018549
0.000523	0.047383	0.010442	-0.003001	0.001361	0.012889	0.044855	0.006653	0.018225	-0.006479	0.02426	0.011483	-0.040979	0.016586	-0.019868	-0.022953	-0.000958	0.029522	-0.038991	0.018145	0.043137	0.018352	0.059474	0.033989	0.00599	-0.067452	0.053045	-0.012063
-0.015785	0.035138	0.0085	-0.013162	-0.014859	0.025962	0.067868	0.002884	0.020419	-0.024244	0.038148	0.012648	0.009806	-0.015439	-0.006422	0.011835	0.035543	0.020241	-0.024555	-0.004794	-0.013844	0.023462	0.025059	0.027734	0.082064	-0.045607	0.069838	-0.02104
0.017036	0.044869	0.022258	-0.020175	-0.051246	0.025094	0.006881	-0.003548	0.005944	-0.017528	-0.001992	0.076991	-0.006618	0.020766	-0.022496	-0.074471	0.004839	0.013826	-0.037867	0.031885	0.00645	0.041806	-0.010242	0.05412	-0.001385	-0.059813	0.031499	-0.018166
-0.051989	0.100271	-0.02211	-0.013968	-0.047048	-0.005323	0.003562	-0.036994	-0.008924	-0.060395	-0.002658	0.001439	0.028579	-0.036211	-0.012587	0.037336	0.021624	0.052739	0.074033	-0.013248	-0.019538	0.027827	-0.032054	0.016516	-0.083349	-0.015933	0.036888	0.01093
-0.027366	0.124492	0.027082	-0.002951	-0.029542	-0.007018	-0.033786	0.034902	0.003245	0.005536	-0.029784	0.085131	0.011736	0.020254	-0.021672	0.012272	0.031172	-0.001395	0.023217	-0.059913	-0.008524	-0.00358	0.053139	-0.000738	0.037517	-0.018837	0.044481	0.075973
-0.028033	0.125162	-0.008227	0.01096	-0.040691	0.004034	-0.046627	0.027559	-0.084463	-0.074938	0.091622	0.059267	-0.040314	0.027271	0.026967	-0.100953	0.053365	0.060634	-0.004663	0.007358	-0.016268	-0.020136	-0.042011	-0.003124	0.006428	-0.031791	-0.0336	0.05124
0.006301	0.055784	0.01364	-0.017838	-0.029339	0.030366	0.004892	-0.012998	0.011017	-0.044831	0.020881	0.007908	-0.004041	-0.011628	-0.003577	0.059838	0.022413	0.071127	0.002641	-0.011375	-0.053064	0.035088	0.044804	0.017636	0.017086	-0.025848	0.029397	-0.031864
-0.012339	0.030535	-0.019365	-0.007551	0.039688	-0.01379	-0.064677	0.038462	0.011453	0.024895	-0.023193	0.076841	0.03485	0.016637	0.033803	-0.014077	-0.029192	-0.002816	-0.020032	0.031607	-0.001899	0.035462	-0.006962	-0.006235	0.001152	-0.088796	-0.012553	-0.027271
0.015664	0.047271	0.010568	0.040867	-0.023277	0.019444	-0.047838	-0.009062	0.004282	-0.090121	-0.006295	-0.025621	0.043741	0.001354	0.053001	0.004262	-0.013402	0.014646	0.037253	-0.048069	0.01203	-0.014902	0.016828	-0.024463	-0.008783	0.046672	0.003464	0.05803
0.043772	0.107332	0.003455	0.027711	-0.017437	-0.009187	0.004532	-0.000793	-0.044961	-0.052115	0.020395	0.036483	-0.066015	0.038322	-0.011871	-0.055982	0.05336	0.011879	0.029169	-0.021944	0.007579	0.024435	-0.017185	-0.023677	0.074318	-0.045457	0.036607	0.033189
-0.024728	0.01367	0.005135	-0.00405	-0.045086	-0.044996	0.03445	0.024503	-0.034046	-0.015758	-0.00858	-0.001793	-0.004013	0.017393	0.032073	0.000331	0.033314	-0.009314	9.82E-05	0.017703	-0.027095	-0.000738	-0.029343	0.008569	0.008257	-0.029392	0.012937	0.030434
-0.025872	0.083091	-0.008578	0.00704	0.024726	0.000447	-0.092274	0.026657	0.008969	-0.038148	0.048593	0.010309	-0.00396	0.010194	0.007646	-0.059267	0.05329	0.030019	0.017011	-0.010065	-0.031686	0.052804	0.009659	0.012491	-0.047325	-0.025421	0.023003	0.088485
0.056444	0.052764	0.034996	0.058723	-0.093539	0.031741	0.072021	0.03773	-0.057536	0.009372	0.036607	0.049081	0.061679	-0.080679	-0.022239	-0.02877	-0.044182	0.047429	0.025924	0.02848	0.031435	0.034201	-0.001239	0.02279	-0.091333	-0.013891	0.012192	0.045744
-0.011025	0.063211	0.019256	0.001904	-0.026545	-0.060628	0.041301	0.012781	0.000569	-0.017318	-0.001072	-0.005518	0.001163	-0.008076	-0.000852	0.004583	0.055479	-0.024674	0.036409	-0.032956	-0.029877	0.035676	-0.002783	-0.003412	0.011293	-0.049897	0.017238	-0.004415
-0.044781	0.115194	-0.013425	-0.034347	0.038582	-0.010362	0.016751	0.014244	0.02742	-0.007971	0.036131	0.008416	-0.041806	-0.003422	0.00433	0.004925	0.018674	0.004121	0.045453	-0.004277	-0.047997	-0.000201	-0.04049	0.011433	0.007936	-0.018741	0.083796	-0.051941
-0.003653	-0.042912	0.030942	0.022327	0.002467	0.010037	0.108596	0.026928	0.025599	0.068218	-0.058213	0.007193	-0.016273	-0.066633	0.022417	-0.039464	-0.000354	-0.031185	-0.035101	0.033816	0.012444	-0.005468	-0.015887	-0.010313	0.012642	0.006648	-0.016446	-0.04454
0.068874	-0.005601	-0.013409	0.027338	0.030566	0.009583	0.033932	0.032312	-0.019138	0.010902	-0.003241	-0.028912	0.040947	-0.048362	-0.016485	-0.018847	0.0377	0.037807	0.038532	-0.032128	0.012074	-0.021966	0.052956	0.007037	0.037156	-0.035967	-0.010116	-0.031169
-0.014998	-0.04261	0.041533	0.008506	0.005618	-0.007203	0.011911	0.039898	-0.00404	-0.018751	0.003272	0.019833	0.026191	-0.025641	-0.029111	0.068299	0.066075	-0.000697	0.103296	-0.006605	-0.056171	0.001674	-0.00274	0.030528	-0.043276	-0.04259	0.061972	0.020934
-0.027462	0.010859	-0.003409	-0.046164	0.050071	0.034046	0.024833	-0.029739	0.025689	-0.024114	-0.041232	0.017575	0.042516	-0.075587	-0.018693	0.036809	-0.006782	-0.001892	0.001712	0.022497	0.03291	0.061391	0.000558	-0.033612	-0.010568	0.003451	-0.056848	-0.028893
0.014458	0.055984	-0.016717	-0.015027	-0.005839	-0.019339	0.018219	-0.023261	-0.083651	-0.009826	0.049513	0.032955	0.027747	0.036437	0.000357	-0.030922	-7.11E-06	0.038372	-0.014657	-0.002652	-0.013652	-0.010548	-0.018346	0.006427	0.067994	0.01102	-0.039511	-0.056267
-0.051366	0.096285	0.040264	0.037584	0.01998	0.013719	-0.015988	-0.035895	0.029737	-0.056048	0.001226	0.071758	-0.02783	0.004694	0.035133	0.00208	0.01292	0.047996	-0.004082	0.013469	-0.011993	0.005089	0.03996	0.013442	-0.003337	-0.001852	0.080917	-0.017113
-0.062097	0.104798	0.011691	0.004429	-0.034061	0.014988	-0.043631	-0.01001	-0.042839	-0.031442	0.04731	0.036755	-0.061662	0.038599	0.041278	-0.051146	0.022752	0.064242	0.014484	-0.044626	0.001732	0.0109	-0.002803	0.006609	-0.01991	0.007549	0.024881	0.038212
-0.03681	0.115386	0.00678	0.035027	-0.053699	0.025966	-0.034983	-0.03442	-0.044123	-0.013542	0.056196	0.069343	-0.059962	0.025709	0.024827	-0.107974	0.039179	0.023371	-0.007798	-0.007572	0.022506	-0.018464	0.006958	0.006469	-0.02811	0.030642	-0.019147	0.027174

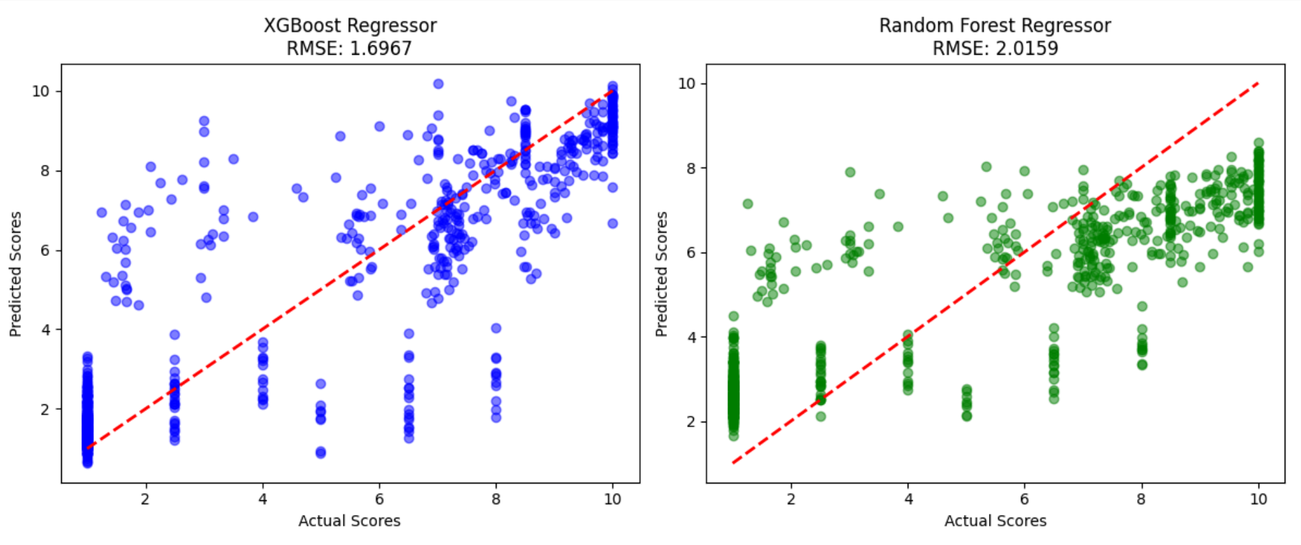
ODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY METHODOLOGY ME

METHODOLOGY - 3

Our Best ML Approach: Embedding-Based Approach

Results after training on dataset with embeddings

Model	MAE	RMSE
XGBoost Regressor	1.209	1.6967
Random Forest Regressor	1.226	2.015



METHODOLOGY - 4

ML Approach: Transformer Approach

- After experimenting with traditional ML models, PCA, and embedding-based pipelines, we implemented a Transformer-based architecture for contextual bias severity prediction.

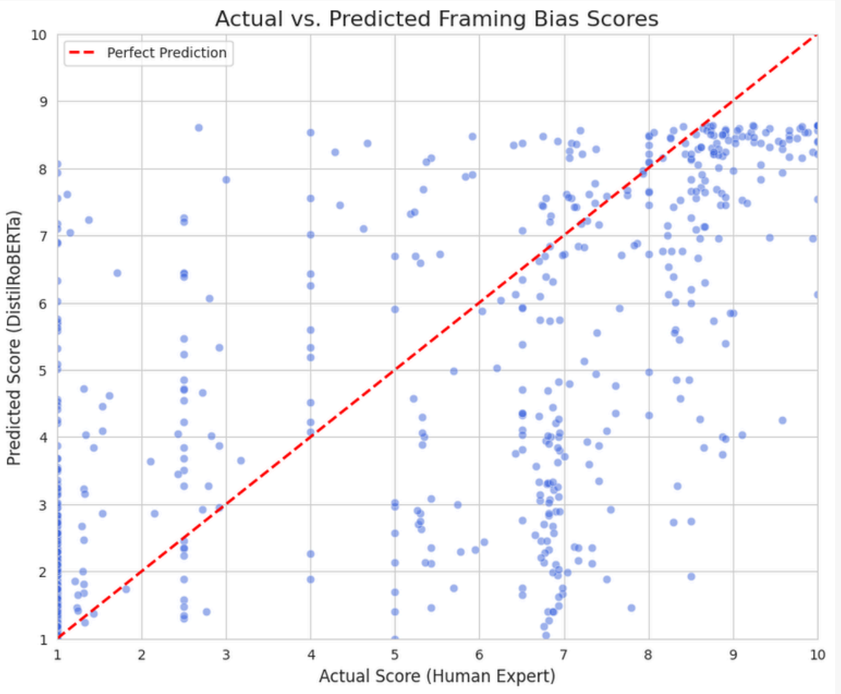
- We selected **DistilRoBERTa** because:
  - Lightweight architecture (~82M parameters)
  - Faster inference and training
  - Strong contextual language understanding
  - Optimized for NLP feature extraction tasks

- Encoder-Only Advantage
  - DistilRoBERTa is an encoder-only transformer, making it highly effective for sentence understanding and content extraction

METHODOLOGY - 4

Transformer Pipeline

- **Tokenization**
  Converted raw text into numerical tensors
  Applied:
  Padding and truncation
  Used DistilRoBERTa tokenizer for standardized transformer input formatting
- **Regression Head Modification**
  Classification → Regression
  Changes Made
  num_labels = 1
- **Supervised Fine-Tuning**
  Used 80/20 train-test split
- **Optimized using:**
  AdamW Optimizer
- Fine-tuned on expert-derived target scores



Metric	Value
MAE	1.57
RMSE	2.23

PERFORMANCE ANALYSIS

Best Performing Model

- **Embedding + XGBoost Pipeline**

	MAE	RMSE	R^2 Score
Metric Value	**1.20**	**1.63**	**0.780**

The hybrid embedding pipeline successfully combined:

- Semantic text embeddings
- Linguistic feature engineering
- Sentiment analysis features
- Bias density indicators

This allowed the model to:

- Capture contextual meaning
- Detect framing intensity
- Preserve computational efficiency
- Generalize better on structured news data

Approach	Model	MAE	RMSE	Key Limitation
TF-IDF Features	Random Forest	2.55	3.1	Poor contextual understanding
TF-IDF Features	XGBoost	2.9	3.18	Sparse high-dimensional vectors
PCA + ML	Random Forest	2.03	2.44	Limited semantic understanding
PCA + ML	XGBoost	1.8	2.23	Context still partially lost
Embeddings + Features	Random Forest	1.226	2.01	Lower semantic generalization
Embeddings + Features	**XGBoost**	**1.2**	**1.63**	**Static embeddings**
Transformer Model	DistilRoBERTa	1.57	2.23	Higher computational cost

APPLICATIONS & DEPLOYMENT

1. **Editorial Pre-screening** Newsrooms can auto-flag highly biased drafts before publication, helping maintain editorial standards at scale.
2. **Algorithmic Transparency** for Platforms Social media ranking systems can demote content with extreme bias scores, promoting healthier information ecosystems.
3. **Media Literacy Tools** for Readers Researchers at CHI 2025 demonstrated that a Media Bias Detector tool was seen by participants as a valuable resource in media literacy classrooms whether at college level or in public-facing settings and could educate users to recognize blind spots in news coverage. ACM Digital Library
4. **Academic Research** Enables large-scale computational journalism studies tracking how framing of the same event evolves across outlets and over time.
5. The solution can be deployed at Plaksha as a lightweight API or browser extension, with potential applications in newsroom editorial screening, social media content ranking, and media literacy education.

THANK YOU

shlok.gupta.ug24@plaksha.edu.in
sartajdeep.s.ug24@plaksha.edu.in
kanav.nanda@plaksha.edu.in