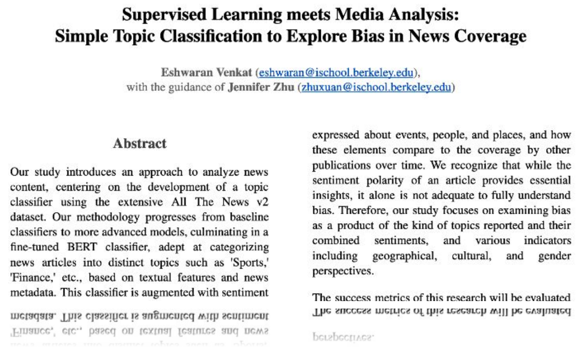
Supervised Learning meets Media Analysis: **Simple Topic Classification to Explore Bias in News Coverage**

Eshwaran Venkat

Jennifer Zhu (Thursday Section),
UC Berkeley MIDS DATACI266

Berkeley
UNIVERSITY OF CALIFORNIA



Supervised Learning meets Media Analysis:
Simple Topic Classification to Explore Bias in News Coverage

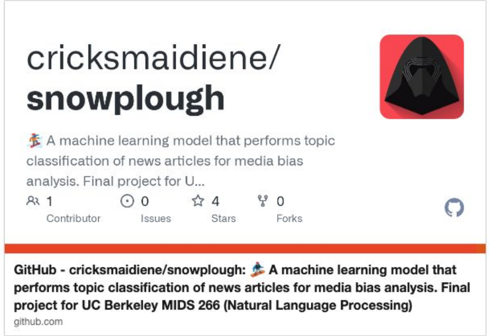
Eshwaran Venkat (eshwaran@ischool.berkeley.edu),
with the guidance of **Jennifer Zhu** (zhuxuan@ischool.berkeley.edu)

Abstract

Our study introduces an approach to analyze news content, centering on the development of a topic classifier using the extensive All The News v2 dataset. Our methodology progresses from baseline classifiers to more advanced models, culminating in a fine-tuned BERT classifier, adept at categorizing news articles into distinct topics such as 'Sports,' 'Finance,' etc., based on textual features and news metadata. This classifier is augmented with sentiment

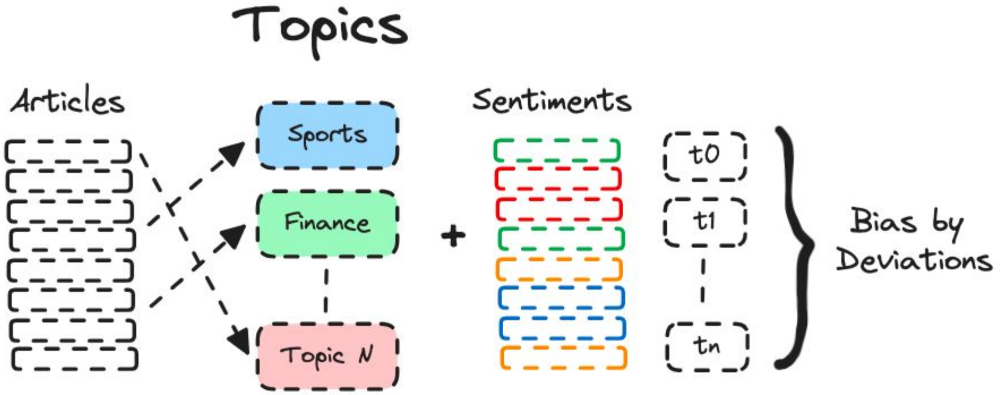
expressed about events, people, and places, and how these elements compare to the coverage by other publications over time. We recognize that while the sentiment polarity of an article provides essential insights, it alone is not adequate to fully understand bias. Therefore, our study focuses on examining bias as a product of the kind of topics reported and their combined sentiments, and various indicators including geographical, cultural, and gender perspectives.

The success metrics of this research will be evaluated



Research Question

How can supervised learning models effectively classify news articles into distinct **topics**, and what does this classification reveal about **media biases** in news coverage?



Data: AllTheNews v2

AllTheNews is a popular dataset of news articles that has two versions. Version 1 & 2.

- Version 2.0 has 2.7 million articles from a number of sources.
- It is a published dataset that is readily downloadable.
- The date range of articles is from January 1, 2016 to April 2, 2020.
- The only metadata available is the article title, publication, section, author, date, and content. We use a subset of these as labels for our classifiers

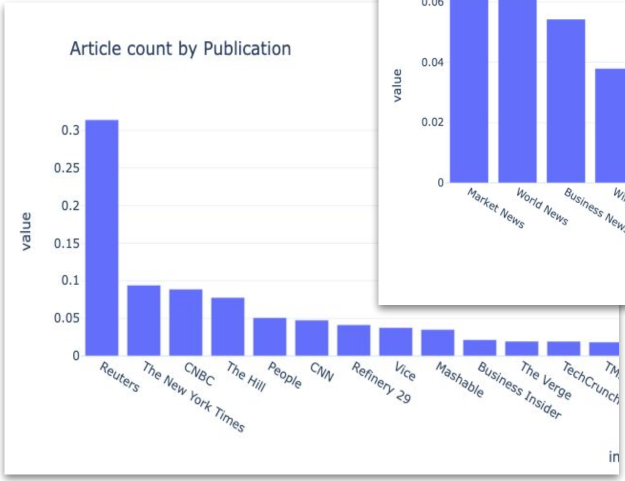
Feature	Description	Count
Date	The date when the article was posted	2.7M
Author	The name of the author who wrote the article	1.6M
Title	The news title of the article	2.7M
Article	The content of the article as text	2.6M
URL	The online hyperlink of the article	2.7M
Section	The newspaper section the article is in	1.7M
Publication	The name of the publication	2.7M

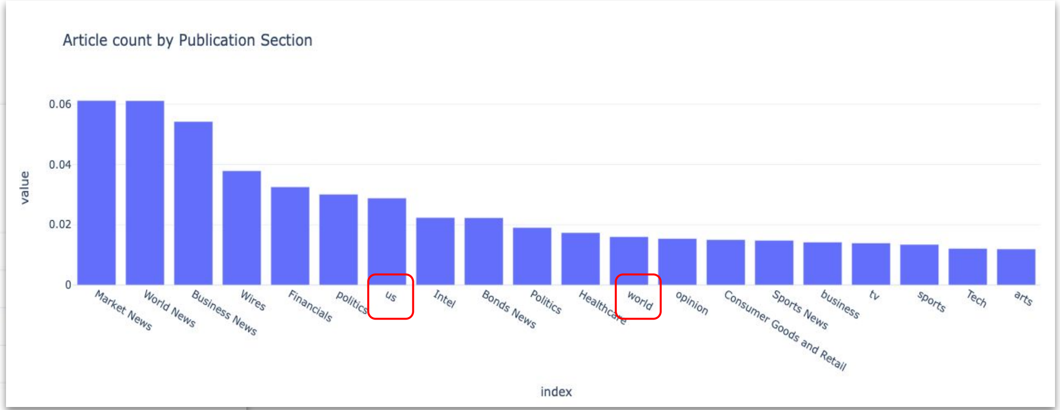
Table 1: Data Dictionary for All The News v2

	date	year	month	day	author	title	article	url	section	publication
1028947	2016-01-01	2016	1	1	Andrea Romano	10 exercises for people with honest New Year's...	In 2016, cut yourself a break. Let this be the...	https://mashable.com/2016/01/01/honest-new-yea...	None	Mashable
1949696	2016-01-01	2016	1	1	Amy Chozick	Hillary Clinton Raised $37 Million in Last 3 M...	Hillary Clinton's presidential campaign raised...	http://www.nytimes.com/politics/first-draft/20...	hillary-clinton-raised-37-million-in-last-3-mo...	The New York Times
590996	2016-01-01	2016	1	1	None	Economic milestones of the year ahead - Daily ...	A NUMBER of economic trends that have been sim...	https://www.economist.com/graphic-detail/2016/...	graphic-detail	Economist
1275364	2016-01-01	2016	1	1	None	Belgium releases three held over New Year atta...	BRUSSELS (Reuters) - Belgian investigators rel...	http://www.reuters.com/article/us-belgium-secu...	World News	Reuters
841225	2016-01-01	2016	1	1	Jodi Guglielmi	Natalie Cole Was 'Full of Gratitude' After Kid...	During Natalie Cole's lifetime, she suffered a...	https://people.com/celebrity/natalie-cole-was-...	celebrity	People

Data: Summary

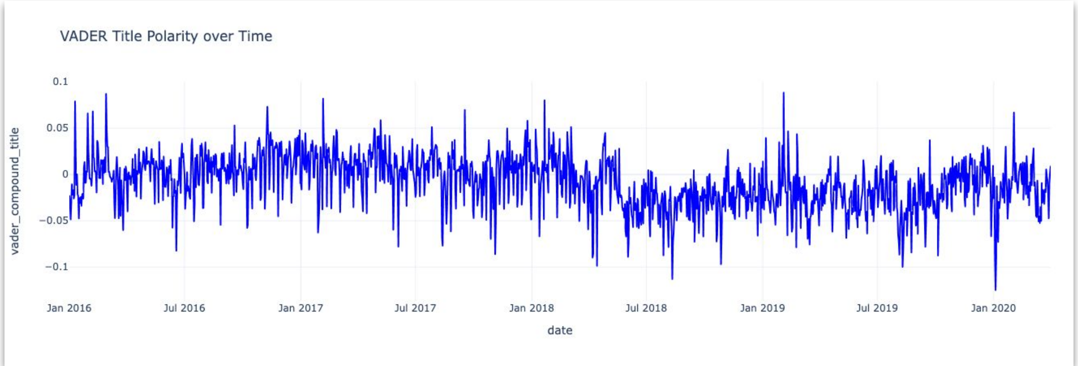
- 30% of articles appear from Reuters
- NYTimes, CNBC & the Hill together make up an additional 25.5%
- 17% of articles come from World News, Business News and Market News sections





Data: Sentiments

- The highly polarized articles are not evident of any bias purely by sentiment
- Refinery 29, CNN, People and NYTimes see an increased presence in polarized articles compared to all articles



	publication	full_ratio	polarized_ratio	polarity_ratio_increase
6	Refinery 29	0.04	0.11	0.06
4	People	0.05	0.08	0.03
5	CNN	0.05	0.08	0.03
1	The New York Times	0.09	0.12	0.02
12	TMZ	0.02	0.03	0.01
16	Washington Post	0.02	0.02	0.01
21	Fox News	0.01	0.02	0.01
24	New Republic	0.00	0.01	0.01

Data: News Sections

- 57% of articles have sections that appear in more than one publication
- 43% of articles have sections that appear only in that publication
- ~200 sections appear across publications, and ~1500 sections are standalone within a pub

Imposter Sections

```
df["section"].value_counts().tail(50)
```

```
ohio-voters-explain-themselves                                          1
retirement-may-be-good-for-you                                          1
a-pitch-for-kindness-outside-trump-tower-in-midtown-manhattan           1
crucible-off-to-strong-start-at-broadway-box-office                     1
ask-well-can-naps-make-up-for-sleep-deficits                            1
why-therapists-should-talk-politics                                     1
word-of-the-day-quiz-prcis                                              1
a-day-of-potential-clarity-amid-a-long-slog                             1
when-is-a-child-too-sick-for-school                                     1
what-were-reading-136                                                   1
morning-agenda-unpacking-a-chinese-companys-spree-on-u-s-hotels         1
new-york-times-anoymous-sources-policy-public-editor                    1
broadways-wicked-soars-past-1-billion-at-the-box-office                 1
1941-yugoslavs-may-delay-nazi-tie                                       1
cancer-family-nancy-borowick-kickstarter-book                           1
march-15-2016-photos-of-the-day                                         1
```

section	publication	
2020 Elections	CNBC	8
	Vice	2
Abortion	Fox News	5
	Vice	3
Advertising	CNBC	45
	Reuters	2
Advice	The Verge	31
	Vice	5
Aerospace & Defense	CNBC	235
	Reuters	12
Afghanistan	Fox News	12
	Reuters	3
	Vice	1
Africa	CNN	2175
	Fox News	19
	Reuters	2
Agriculture	CNBC	100
	Reuters	148
Airlines	CNBC	1199
	Fox News	187

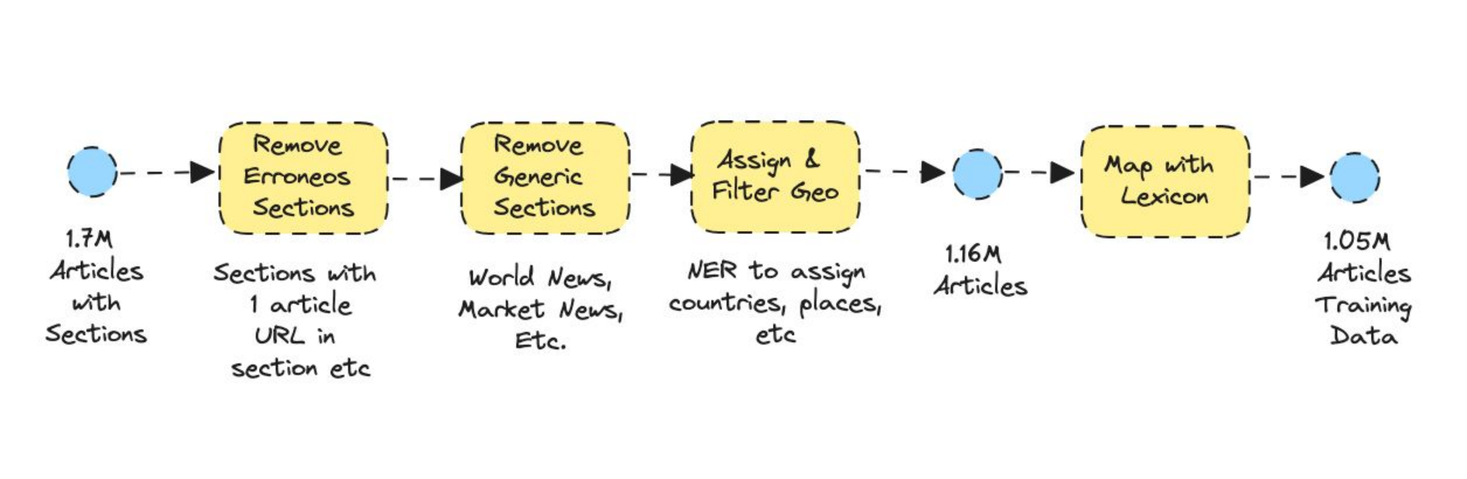
Need Coalesced / Normalized Section Names:

- Reduce target classes
- Minimize Overlap of Topics
- Minimize Overlaps b/w Publications

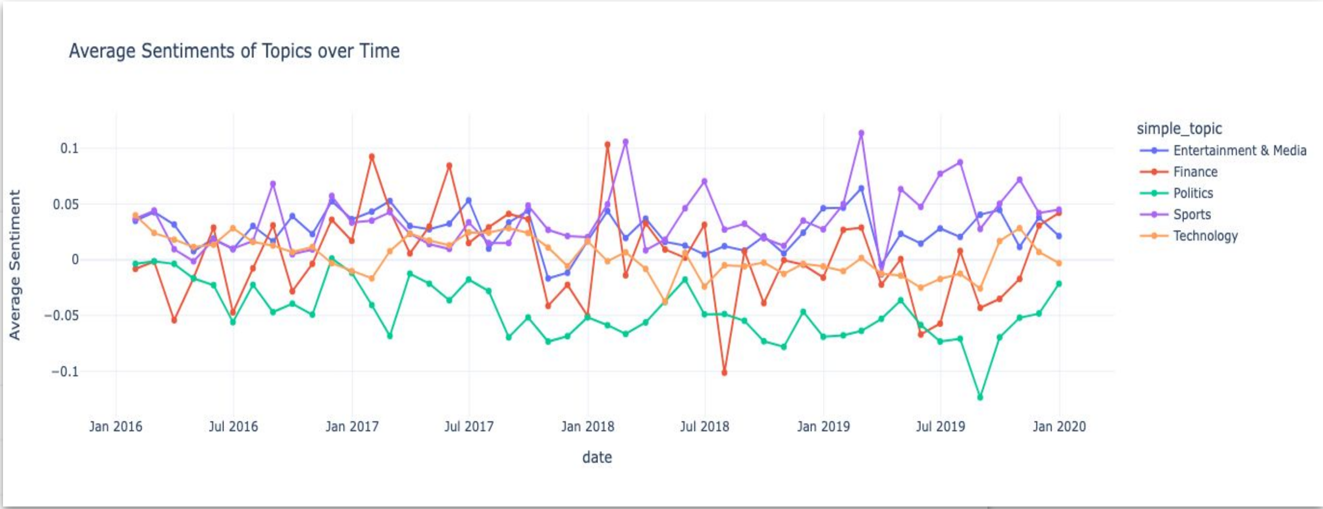
	section	article_count	section_clean	simple_section_topics	simple_topic
107	Olympics News	1919	olympics	[Sports]	Sports
831	E-Sports	10	sports	[Sports]	Sports
1415	National Basketball Association	2	national basketball association	[Sports]	Sports
4	Sports	35132	sports	[Sports]	Sports
339	Olympics Rio	99	olympics rio	[Sports]	Sports

Feature Engg: Topics

- Utilizes a topic lexicon to map existing news sections to topics.
- The lexicon encompasses around 500 strings across less than 20 news topics.
- Lexicon considers **lemmatization**, **singular** and **plural** forms, differentiates between substrings (to avoid cases like 'Car' being confused with 'Carpet'), and recognizes **hyphenated** words. Additionally, it is fine-tuned to handle **short strings with precision**, ensuring that brief section names like 'War' or 'Bus' are accurately categorized without being conflated with longer, unrelated terms



Feature Engg: Topic Exploration



Baseline Models

- Selected Baseline Models:
  a. Logistic Regression
  b. Multinomial Naive Bayes
  c. Random Forest Classifier
- Feature Processing:
  a. Applied TF-IDF vectorization for feature extraction from text data
  b. Separate vectorization for titles and articles
- Performance Evaluation:
  a. Focused on weighted F1 Scores for model comparison

		Title	Article (1Y)
Models	**Logistic Regression**	64%	78%
	Random Forest	66%	76%
	Naive Bayes	60%	69%

	precision	recall	f1-score	support
Art & Culture	0.68	0.46	0.55	11023
Commercial Business	0.56	0.60	0.58	26266
Education	0.72	0.44	0.54	840
Entertainment & Media	0.59	0.78	0.67	32150
Environment	0.65	0.47	0.54	4388
Finance	0.72	0.78	0.75	33731
Health	0.68	0.57	0.62	14515
Humanitarian	0.00	0.00	0.00	8
Industrial Business	0.50	0.13	0.20	1436
Journalism & Opinion	0.93	0.82	0.87	8805
Legal & Safety	0.65	0.55	0.60	3215
Living	0.62	0.37	0.47	7554
Military	0.31	0.03	0.05	348
Politics	0.72	0.80	0.76	21267
Science	0.51	0.37	0.43	2869
Sports	0.77	0.75	0.76	12991
Technology	0.67	0.61	0.64	20067
Transportation	0.41	0.21	0.28	1137
Weather	0.34	0.13	0.19	542
accuracy			0.67	203152
macro avg	0.58	0.47	0.50	203152
weighted avg	0.67	0.67	0.66	203152

Topic Classification from Title - Random Forest Regressor

Class imbalance study showed no significant changes

Neural Nets

- **Bi-Directional LSTM**:
  a. **Architecture**: Utilizes a Bidirectional LSTM with 128 and 64 units, coupled with a 128-unit dense layer.
  b. **Rationale**: Bi-LSTM layers capture both forward and backward context in text, essential for multifaceted narratives in news stories. The dense layer aids in balancing complexity

- **CNN**:
  a. **Architecture**: Begins with an embedding layer converting text into 64-dimensional vectors. Employs `Conv1D` layers with 64 and 128 filters, followed by `GlobalMaxPooling1D` and a 64-unit dense layer with ReLU activation.
  b. **Rationale**: The embedding layer provides a compact representation suitable for CNN's pattern recognition approach. Conv1D layers capture local text features effectively, and the pooling layer reduces dimensionality, emphasizing significant features

Models were chosen as they're used to **Fine-Tune BERT**

Model Results better than baseline at around **70-72% F1** for news titles to topics (5-7 pp increase)

Fine-Tuned BERT Models

BERT Simple

- Built on `bert-base-uncased` pre-trained model
- Design Rationale:
    a. Aimed to achieve lower model complexity
    b. A **dropout** layer with a rate of 0.3 follows the BERT output to reduce overfitting

Models	Title	Article (1Y)
BERT Simple	77%	84%
BERT Complex	78%	86%

Table 5: Fine-tuned Bert-Base-Uncased Model Results - Weighted F1 Scores on 20% test data when classifying topics.

BERT Complex

- **Tokenizer & BERT Model**:
    a. Begins with a BERT tokenizer that preprocesses input text
    b. Utilizes the "bert-base-uncased" model for initial text representation
- Additional Layers:
    a. **Bidirectional layer**: Enhances the model's ability to understand context by considering both previous and subsequent tokens
    b. **Global Average Pooling 1D**: Reduces dimensionality and summarizes key features from the bidirectional layer
- **Dense layers with Layer Normalization**: Provide further non-linear transformation of features and aid in final classification avoiding overfitting

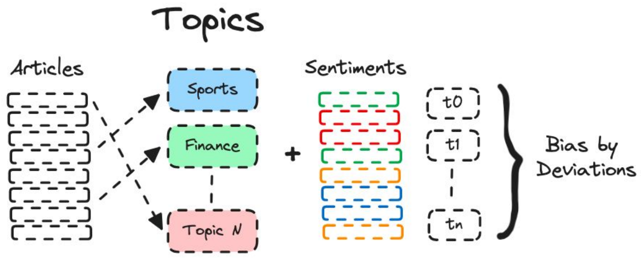
BERT Complex

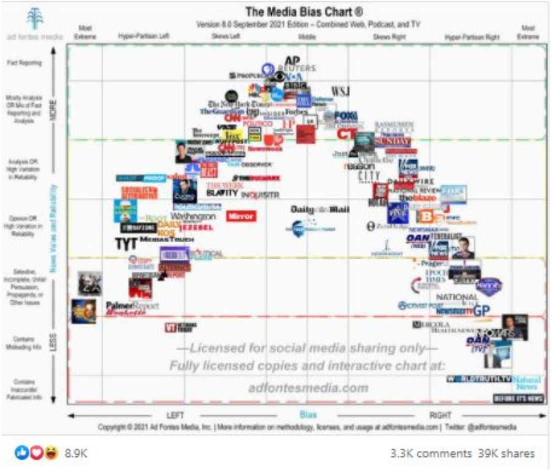
Trainable Params: 110M

Layer (type)	Output Shape	Param #
input_ids (InputLayer)	[(None, 64)]	0
tf_bert_model (TFBertModel)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 64, 768), pooler_output=(None, 768)	109,482,240
bidirectional (Bidirectional)	(None, 64, 256)	918,528
global_average_pooling1d (GlobalAveragePooling1D)	(None, 256)	0
dropout_37 (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32,896
layer_normalization (LayerNormalization)	(None, 128)	256
dropout_38 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 19)	2,451

Simple Bias Analysis

We define a **bias (deviation) score** as a metric that combines two factors: the proportion of articles that a publication contributes to a specific topic in a given month (**published ratio**) and the sentiment of those articles compared to the overall sentiment for that topic (**sentiment ratio**) for the same month. This score has been averaged for each publication-topic combination over each month.





Potential Improvements

- More complex bias analysis
- Improving the reliability of the topic lexicon by consensus, and support more topics
- Using metrics other than Accuracy, Precision, Recall, F1 and Weighted F1 (like AUC)
- Use more than 1y article data (since it's better than title)

Shoutout to Jennifer!

Q&A