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 analysis and other indicators for a supplemental exploration into media bias, aiming to delineate its various manifestations. The core of our research lies in the robust topic classification, with media bias analysis providing additional insights.

We've made the code, notebooks, models and newly generated (topic classified) dataset publicly available. The newly created dataset is listed as All The News v2.1 on Kaggle and a fine-tuned BERT classifier for the same is also made available online.

Code: github.com/cricksmaidiene/snowplough
Project Page: cricksmaidiene.github.io/snowplough

Introduction

Bias in a news article refers to the presence of a certain slant or inclination in the way information is presented. It means that the content is delivered with a partial perspective, favoring one side or viewpoint over others. This can manifest in various forms, such as political, cultural, commercial or geographical bias, among others. The rise of digital media has magnified the influence of news reporting on public perception, bringing the issue of media bias to the forefront of societal concern.

In our research, we operate under the premise that analyzing the bias of a news article in isolation is insufficient, as bias is inherently contextual, influenced by the topics covered, the sentiments 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 on two fronts: First, the performance of the topic classifier, measured using machine learning metrics like accuracy and weighted F1 scores, and across the individual trained labels. Second, the effectiveness of the descriptive bias analysis, mostly by subjectively comparing with existing sources like AllSides.com, MBFC and otherwise through expert opinion. While the former provides a quantitative assessment of the classifier's ability to categorize news topics accurately, the latter, being more subjective, will focus on the depth and relevance of the insights generated.

Literature Review

D'Alonzo and Tegmark [1] develop a method to infer news article sources using phrase frequency, creating a "bias" landscape. Raza et al. [2] introduce Nbias, an NLP framework for bias identification, combining transformer-based models and diverse data sets. Spinde et al. [3] employ a Transformer-based MTL approach with DistilBERT for bias detection in news, focusing on word choice bias.

Contemporary research on news topic classification focuses on multimodal methods, combining text and image data for enhanced classification. For instance, the N24News dataset [4] incorporates both text and image information. Recent studies also explore the application of deep learning techniques in text classification, demonstrating the effectiveness of

models like CNN, LSTM, and BERT Models in capturing the relationship between words and news categories [5-6]. A number of sentiment analyzers have been used with neural nets as well as fine-tuned transformers to classify news sentiments as well [7-11].

Dataset

Datasets that were otherwise considered but not incorporated included Multiple fake news classification datasets, Media Bias Fact Check Database (MBFC), AdFontes Media Bias Data and Huffington Post News Categories. Their exclusion was due to non-compliance with our criteria concerning volume and label distribution, alongside the presence of ample existing research for certain datasets and prohibitive resource demands for extraction of others.

The dataset selected for our research is the expanded version of the "AllTheNews" dataset, known as All The News v2. This comprehensive dataset includes approximately 2.7 million news articles sourced from 27 American news outlets, covering a timeline from 2016 to mid-2020. This dataset has been designed primarily for research purposes and its smaller predecessor is widely recognized for its use in Kaggle challenges. Our choice to utilize this dataset was informed by a couple of factors:

- **Extensive & Verifiable Corpus**: Represents a significant random sample of real and recent news events, with hyperlinks to the original articles. This enhances the authenticity and verifiability of the data and also provides sufficient volume for training large models.
- **Labels**: Presence of a large number of author, publication and newspaper section values which can be used as training labels.

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

Note from Table 1 that Section and Author have about 40% null values. Our target is to fill up values for "Section" (through the topic classifier) in order to conduct our bias analysis.

30% of articles appear from the Reuters publication. NYTimes, CNBC & the Hill together make up an additional 25.5%. Publications like Refinery 29, People and CNN see at least a 3 percentage point increase in the total ratio of articles when considering highly polarized articles by simple sentiment scoring, compared to their ratio in the entire dataset.

The section of an article highlights which section of the newspaper it appears in. 17% of articles come from World News, Business News and Market News sections which are too general to be classified into specific topics. 57% of articles have sections that appear in more than one publication, and publications may have intersecting topics but different section names (ex. Tech by VICE, Technology News). Distinct sections per month over time are fairly uniform and can be geographical (ex. Australia, Afghanistan), or topical like the 2020 elections.

Data for the year 2020 only exists till April 2 2020, and due to the highly polarizing nature of the first half of that year, only data till Dec 2019 is considered. Daily article count is approximately uniformly distributed by month over all years.

Authors like Dave Quinn, Alexia Fernandez & Stephanie Petit have written more than 4000 articles each. Some authors like Alex Shephard and Jeet Heer have written more than 10% of all articles seen in their publications, with many others crossing 5% of the total articles in their respective publications. This is of note if our dataset is a representative sample of the population.

VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiments were assigned to all article titles together with textblob sentiments. Both models showed highly comparable results over time, and the average sentiment centered between -0.05 to +0.05. VADER is preferable for our analysis since it

provides probabilities of positive, neutral and negative sentiments.

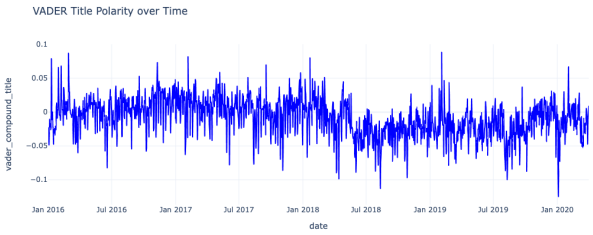


Fig 1: VADER model assigned sentiment scores for all articles over time. Values range as -1 being negative, 0 being neutral and 1 being positive

Topic Labeling

We intend on using the newspaper section names as topic labels in order to train the topic classifier and fine-tune transformer models. We chose section names because they inherently represent well-defined, established categories that align closely with the general public's understanding of news topics, leverage the editorial expertise of the publications, and eliminate the need for manual labeling.

The raw section names cannot serve as direct topic labels due to their abundance, which would lead to an impractical number of target classes for supervised learning. Additionally, these names vary in naming conventions across publications, making them non-mutually exclusive in terms of topic semantics, as highlighted in the dataset summary.

We first remove erroneous section names, such as those that have 1 article per section, or are otherwise non-representative of section names. We also remove highly generic sections that carry a diverse set of topics like World News, Market News, Wires, etc. Geographical entities are then assigned to sections using Named Entity Recognition (NER). This helps tag and filter for country specific news, which itself may have a distribution of diverse news topics that we do not have sub-labels for. Around 14% of sections are mapped to geographies, and that constitutes 11.3% of all articles. After filtering out geographic entities as well, we arrive at 1.16M articles (from an initial value of 1.7M) with sections that can be mapped to topics with our approach.

We have created a simple yet direct approach of mapping sections to topics using a section-to-topic lexicon. The lexicon encompasses around 500 strings across less than 20 news topics. The emphasis was on mapping major sections, representing a substantial ratio of all articles, rather than a complete section-to-topic correlation.

Topic	Description	Sections Mapped
Finance	News around the Fed, Banking, Stocks, Earnings, etc.	108
Technology	Tech, Gadgets, AI, Computers, Electronics, etc.	81
Entertainment & Media	Movies, TV Shows, Magazines, etc.	81
Art & Culture	Museums, Fashion, Books, Photography, etc.	60
Commercial Business	Retail, Consumer Goods, Consulting, etc.	54
Politics	Elections, Campaigns, Political Parties, etc.	47
Living	Pets, Dining, Home, Apparel, Accessories, Babies etc.	47
Health	Healthcare, Pharmaceuticals, Medical	46
Sports	NHL, NFL, MLB, Soccer, Olympics, etc.	27
Legal & Safety	Legal news, Crime, Homicide, Police, etc.	26
Environment	Environment, Sustainability, Clean Energy	21
Science	Physics, Mathematics, Space, Animals, Geology, etc.	22
Journalism & Opinion	Op-Ed, Letters to the Editor, etc.	20
Military	Drones, War, Tactics, Air Force, Navy, etc.	15
Industrial Business	Farming, Automotive, Oil & Gas, etc.	17
Education	Learning, Pedagogy, Exams, etc.	16
Transportation	Airlines, Airports, Rail, etc.	11
Weather	Weather, Climate, Hurricanes, etc.	11
Humanitarian	Altruism, Human rights, Charities, Fundraisers, etc.	5

Table 2: Topics assigned through the Lexicon and what their coverage looks like

The core of this process is a specialized function designed to intelligently match section names to topics based on the lexicon. This function is adept at recognizing and interpreting various string patterns within section names. It takes into account not just exact matches but also variations and nuances in the text. For instance, it 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.

The lexicon maps ~715 sections to topics out of ~1500 total sections. The mapping converged here since the remaining sections have a low number of articles and generally either capture more niche ideas, or would not contribute to the existing labels significantly for training models. 1.01M articles out of 1.16M were successfully mapped to topics from their sections as a result. Rigorous manual examination and tests were conducted on the mapped sections with plenty of iterations and updates. Care was taken to preserve the original section's semantics. The distribution of publications before-and-after the mapping & filtering did not see a statistically significant change in percentage points. This method, while not ideally scalable for very large models, is crucial in a field where categories and terminologies are constantly changing. We further acknowledge that a significant portion of results can be affected by the correctness and consensus of these mappings.

Baseline Models

The mapped topics constituted a total of ~1 million articles, and titles and articles were pre processed through lemmatization, stopwords removal and other techniques to represent fairly clean vectors for training. All baseline models were trained using TF-IDF vectorization that turned the cleaned input features of titles and articles independently into separate feature vectors. TF-IDF increases with the rarity of a word and its frequency in a document, making it a powerful tool for extracting relevant features. Three main models were used for comparison: Logistic Regression, Multinomial Naive Bayes, and Random Forest Classifiers. SVM was not used as the training time was too large for the entire dataset, and a sample baseline test with other models indicated that its cost was not worth pursuing for the comparable results it provides. The 3 classifiers were trained on the data in two broad ways. Those that predict the news topics from the title of the news article, and those that predict from the article content itself. Since the size of articles is quite large, only a single year's article data was used for the baseline, as opposed to the news title which had a train-test split on the complete dataset. Different maximum word counts for the TF-IDF vectorizer were tested, and the best results are made available in the table below.

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

Table 3: Baseline Model Results - Weighted F1 Scores on 20% test data when classifying topics from titles and articles. 5k TF-IDF vectorizer used for title, and 10k for articles. Articles baselines are from 1 year's worth of news data only due to size limitations.

The "military" topic consistently had both poor support (representation of articles) and F1-scores. Topics like Sports, Journalism & Opinion, Finance & Entertainment consistently scored higher at more than 80%, and topics with a support less than 1k articles had accuracies no greater than 60%. To investigate if potential class imbalance could be problematic, we resampled data into different bins of uniformly distributed topics by article count, and also used thresholds to remove low support topics like "Humanitarian" & "Military". We observed that the baseline results did not change significantly on these tests, and kept the original results.

Deep Learning Models

In the next section, we fine-tune a BERT model using additional layers from neural nets like RNNs and CNNs. This section tries these architectures independently on the data so as to ascertain any deltas between the baseline results, and the future results when fine-tuned with BERT.

Bi-Directional LSTM

An embedding layer is first used to convert textual input into dense vectors. Bi-D-LSTMs are preferred for their ability to process both forward and backward context in text, a key requirement for the multifaceted narratives in news stories. The model incorporates two LSTM layers, which were experimented on with different layer sizes. A dense layer with ReLU activation is employed for its efficient handling of non-linear data transformations, and finally to prevent overfitting, a dropout layer is included.

The model's best performing hyperparameter configurations are as follows: maximum sequence length is set to 32 words, ensuring efficient processing while capturing relevant context. The vocabulary is limited to 10,000 words, an embedding dimension at 128 provides a detailed representation

of words in a multidimensional space. A learning rate of 0.0001, with 5 epochs and a batch size of 32, is chosen for consistent learning. LSTM layers with 128 and 64 units, and a dense layer of 128 units, are selected to balance model complexity and overfitting risk.

CNN

The model starts with an embedding layer, converting text into 64-dimensional vectors, a more compact representation compared to the previous model, suitable for the CNN's pattern recognition approach. The CNN uses Conv1D layers with filter sizes of 64 and 128, chosen for their capability to capture local patterns and features in the text. The kernel size of 5 allows the convolutional layer to consider small groups of words together, aiding in understanding local context. This is followed by a GlobalMaxPooling1D layer, which reduces dimensionality and extracts the most significant features from the convolutional layer outputs. A dense layer with 64 units and ReLU activation is used for decision-making, similar to the previous model. To mitigate overfitting, a dropout layer with a rate of 0.5 is included. The model operates with a maximum sequence length of 20 words, focusing on more concise text segments, reflecting the neural net's strength in extracting local and relevant features. A vocabulary size of 10,000 is maintained, along with a learning rate of 0.0001, batch size of 64, and 5 training epochs.

Both models perform better than the baseline, and the neural networks are able to capture more nuances in the text data independently.

Models	Title
Bi-Directional LSTM	72%
CNN	69%

Table 4: Neural Network Model Results - Weighted F1 Scores on 20% test data when classifying topics from titles only. Only best model architectures shown.

Fine-Tuned BERT

Two kinds of BERT models were fine-tuned, and both were considered from the "bert-base-uncased" pre-trained model. BERT Simple, which has lower model complexity, and BERT Complex, which has denser layers. Find model architecture summaries in the appendix.

BERT Simple

The architecture incorporates a maximum token length of 64. A dropout layer with a 0.3 rate follows the BERT output, aiming to reduce overfitting by randomly deactivating a portion of neurons during training, a necessary step given the complexity of the model. The final layer is a dense softmax activation layer, structured for multi-class classification by producing probabilities for each class, suitable for discrete news topics.

BERT Complex

The architecture begins with a BERT tokenizer that preprocesses input text, followed by a BERT model layer which generates contextual embeddings from the input. After the BERT layer, a Bidirectional LSTM (Bi-LSTM) with 128 units is applied. Subsequently, a Global Average Pooling 1D layer is used to condense the output of the Bi-LSTM layer, reducing its dimensionality and summarizing its features. This is followed by a dense layer with 128 units and a ReLU activation function, providing further non-linear transformation of the features. A Layer Normalization is then applied,

To mitigate the risk of overfitting, two dropout layers are included: the first with a rate of 0.3 after the pooling layer and the second with a rate of 0.5 after the dense layer. The final layer of the model is another dense layer with a softmax activation function. This layer outputs the probabilities for each class, corresponding to different news topics.

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 Simple does surprisingly well enough, and the intricacies captured in BERT complex models don't necessarily translate into significantly better results.

News Media Bias Analysis

A longitudinal approach will be taken to understand macro trends for topics, sentiments and indicators. We then consider a cross-sectional study across specific time intervals for bias indicators. We first use the topic classifier, using a threshold of labeling confidence to label as much of the remaining data in AllTheNews v2 as possible. For instance, if the classifier labels an article as 90% belonging to Sports, and even if that article did not have a section to begin with, we label it with our new-formed topics.

Once this is complete, we also VADER sentiments to all articles, such that we can begin to estimate potential biases. We also have the geographic entities from topic labeling which we can use to identify if there are geographic biases.

We can consider our preliminary analysis through a set of data-driven questions.

1. Topics & Sentiment Trends

Are certain topics reported with consistent sentiment compared to others, and does this depend on the publication?

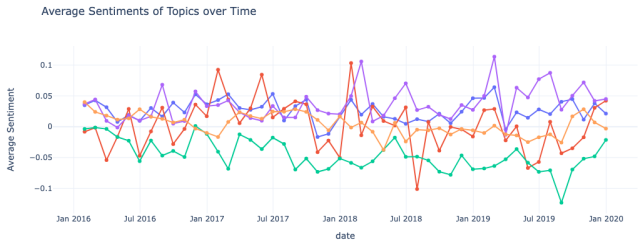


Fig 2: Sentiment trends of assigned topics using the classifier together with existing Data. Topics: Entertainment & Media, Finance, Politics, Sports, Technology are represented as Blue, Red, Green, Purple and Orange respectively. This shows a sample of topics and their complete trends over time, both of which are significant factors for considering bias.

Negative sentiments for "Legal & Safety" sections should not be considered, since they deal with sub-topics like crime, which are generally reported with an inclination. The otherwise average sentiment of topics across publications that stood out from generally neutral thresholds are as follows.

topic	publication	Average sentiment
Commercial Business	Gizmodo	-0.52
Education	Economist	0.26
	Gizmodo	0.48
Environment	CNN	-0.27
	The Verge	-0.51
Humanitarian	Vice	0.49
Legal & Safety	CNN	-0.40
	Fox News	-0.37
	People	-0.40
	Vice	-0.33
Living	Vice	-0.33
Military	Gizmodo	-0.28
	Vice	-0.38
Politics	Washington Post	-0.39
Sports	CNN	-0.36
	Gizmodo	-0.53
Transportation	Fox News	-0.25

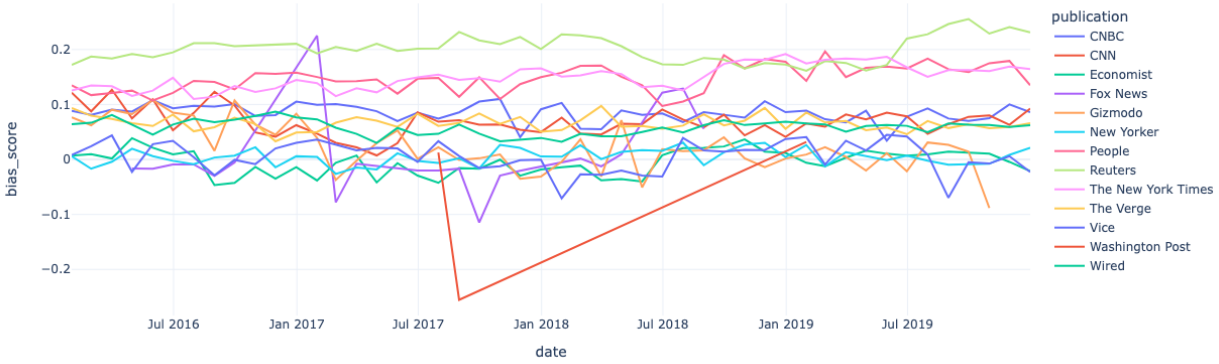
Table 6: Topics across publications by average sentiment. -1 indicates negative, 0 as neutral and 1 as positive.

The average sentiment over time across all publications however, is generally neutral meaning that more volatile articles, topics and publications are drowned out with more number of neutral articles.

2. Scoring & Deviation

We define a bias 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.

A score, such as 0.2, reflects the extent to which a publication's coverage of topics deviates from the average in terms of volume and sentiment. The score is normalized, typically ranging from -1 to 1, where 0 indicates no bias (i.e, The publication's coverage is in line with the average), positive values indicate a positive bias (either more articles or more positive sentiment than average), and negative values indicate



a negative bias (either fewer articles or more negative sentiment than average).

If a publication frequently scores highly positive (e.g., around 0.2 or higher), it might be focusing more on certain topics compared to others or portraying those topics in a more positive light than the average. Conversely, consistently negative scores might suggest a lack of coverage or negative sentiment towards certain topics compared to the average. If a publication's bias score varies greatly over time, this might indicate a change in editorial focus, response to specific events, or shifts in audience or market strategy. Consistent scores might indicate steady editorial policies or perspectives

Appendix A: Tools & Infrastructure

Due to the large size of the dataset, it's worth taking note of the platforms and infrastructure required to reproduce results if necessary. All descriptive analysis, data engineering, processing and baseline modeling was run within Python environment-based Databricks notebooks on CPU backed single-node clusters. Spark was not required, and the main choice for Databricks here was to allow variable sized clusters based on requirements at different project stages. No Databricks-specific commands or dependencies exist, and the **notebooks are agnostic and can be run directly on Jupyter or Google Colab as well**, provided that the Python requirements are met, and the requisite hardware is available. A custom Delta Lake (an open source file format on top of apache parquet) handler to store data locally in the file system or on AWS S3 was used, in order to manage memory better for the size of All The News v2. The neural network based classifiers were trained on P-class and G-class instance-type GPUs made available through AWS & Databricks. Mlflow was used to track and save experimental results for trial and error of hyperparameter tuning.

Appendix B: BERT Model Summaries

BERT Simple

Trainable Params: 109M

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
dropout_37 (Dropout)	(None, 768)	0
dense (Dense)	(None, 18)	13,842

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

References

1. S. D'Alonzo and M. Tegmark, "Machine-Learning media bias,"*arXiv preprint arXiv:2109.00024, 2021.* [Online]. Available: https://ar5iv.org/abs/2109.00024.

2. S. Raza et al., "Nbias: A Natural Language Processing Framework for Bias Identification in Text," arXiv *preprint arXiv:2308.01681, 2023. [Online].* Available: https://ar5iv.org/pdf/2308.01681.pdf.

3. T. Spinde et al., "Exploiting Transformer-based Multitask Learning for the Detection of Media Bias in News Articles," *arXiv preprint arXiv:2211.03491,* 2022. [Online]. Available: https://ar5iv.org/pdf/2211.03491.pdf.

4. Wang, Shan, Zhang, Yang., N24News: A New Dataset for Multimodal News Classification." *arXiv preprint arXiv:2108.13327, 2021. [Online].* Available: https://ar5iv.org/abs/2108.13327.

5. Dairui Liu, Derek Greene, and Ruihai Dong. 2022. A Novel Perspective to Look At Attention: Bi-level Attention-based Explainable Topic Modeling for News Classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2280–2290, Dublin, Ireland. Association for Computational Linguistics.

6. Chuhan Wu, Fangzhao Wu, Yang Yu . 2021. NewsBERT: Distilling Pre-trained Language Model for Intelligent News Application. In Findings of the Association for Computational Linguistics: ACL 2021.

7. Vicente, B., Lillo, J., Sáez-Trumper, D., (2021). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. PLOS ONE.

8. Wang, Y., Kang, S., (2022). Text Sentiment Analysis Based on Transformer and Augmentation. Frontiers in Psychology.

9. Arora, P., Srinivasan, R., (2021). Comprehensive Review on Transformers Models For Text Classification. In 2021 4th International Conference on Intelligent Sustainable Systems (ICISS). IEEE Xplore.

10. Liu, J., Chen, Y., (2020). Fine-Tuned Transformer Model for Sentiment Analysis. In Advances in Intelligent Systems and Computing. Springer Link.

11. M. Yu, M. Li, (2021). Sentiment Analysis Based on Bert and Transformer. In Advances in Intelligent Systems and Computing. Springer Link.