Assessing the evolution of public sentiment towards artificial intelligence (AI) over time in news media and their key themes

Pranav Thorat

NYU Courant Institute of Mathematical Sciences

pat9991@nyu.edu

Motivation:

The AI landscape itself is evolving at break-neck speed. Breakthroughs such as large language models, autonomous vehicles, and facial-recognition systems introduce fresh ethical questions and spur new media narratives. We still lack a data-driven account of (i) how journalistic tone toward AI has changed across these inflection points, (ii) which themes dominate coverage, and (iii) how sentiment and theme interact. Leveraging large-scale text mining, modern transformer-based sentiment classifiers, and probabilistic topic models promises a more granular, longitudinal picture than surveys or manual content analyses can provide.

Research Questions:

RQ1 – Temporal Sentiment: How has overall sentiment toward AI in major English-language news outlets evolved from 2012 to 2025?

RQ2 – Thematic Landscape: What latent topics recur most frequently in AI-related news coverage, and how have their relative prevalences shifted over the same period?

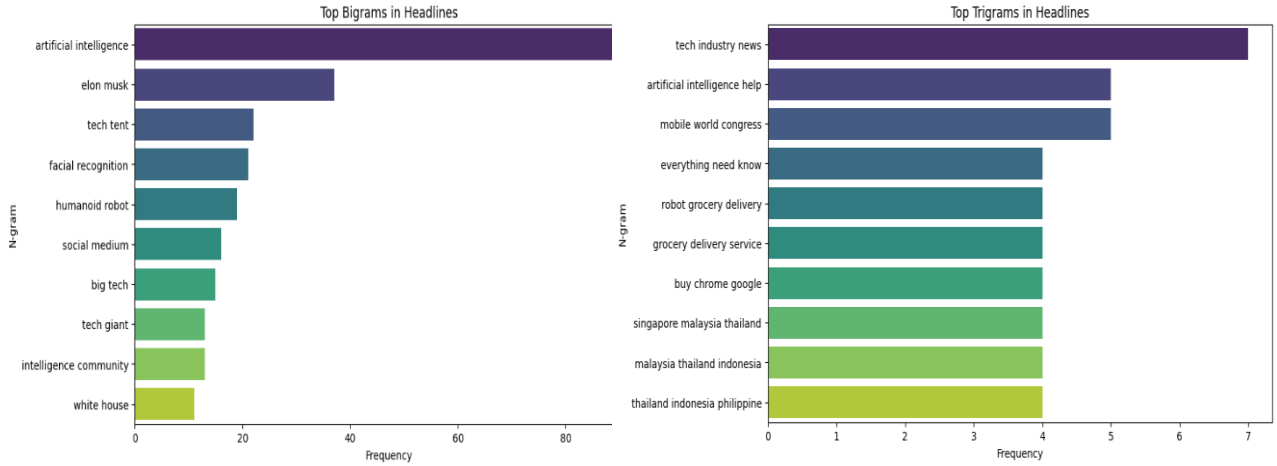
RQ3 – Sentiment–Topic Coupling: How does sentiment polarity vary across the identified themes (e.g., workforce automation vs. generative AI), and which topics attract consistently positive or negative framing?

RQ4 – Methodological Performance: To what extent do state-of-the-art transformer models (BERT) outperform classical baselines in classifying news sentiment, thereby enabling reliable large-scale monitoring of media tone?

Each research question hinges on how journalistic narratives are articulated, text is the primary evidence. Sentiment trends (RQ1) manifest in word choice and framing; latent themes (RQ2) emerge from recurring term co-occurrences; and sentiment–topic interactions (RQ3) can only be observed within the linguistic context that links evaluative language to specific AI subtopics. While auxiliary signals such as images or social-media engagement could enrich the analysis, they cannot substitute for the lexical and semantic content that directly encodes tone and subject matter. Therefore, large-scale text mining offers the most precise and comprehensive lens for answering the study's core questions.

Dataset Description

The project used a sentiment-annotated news dataset, expanded via a custom web scraping pipeline to collect ~5,000 AI-related articles from diverse sources, including general news, tech, business, and academic outlets. Each article was processed to extract key fields (ID, source, headline, date, and lead content). Articles lacking sentiment labels were annotated using a pre-trained classifier to ensure full sentiment coverage.



Text Preprocessing and Outlier Handling

The raw textual data sourced from various news articles was systematically cleaned and structured to prepare it for robust sentiment and thematic analysis. Initially, data loading accounted for potential encoding discrepancies (such as UTF-8, ISO-8859-1, Latin1, and CP1252). The dataset was then refined by removing duplicate entries based on unique text identifiers like headlines and lead paragraphs. Entries with missing values in essential metadata

fields—such as serial numbers, source outlet names, headlines, publication dates, and lead summaries—were either imputed or eliminated to ensure consistency. Dates were standardized uniformly to the DD-MMM-YY format.

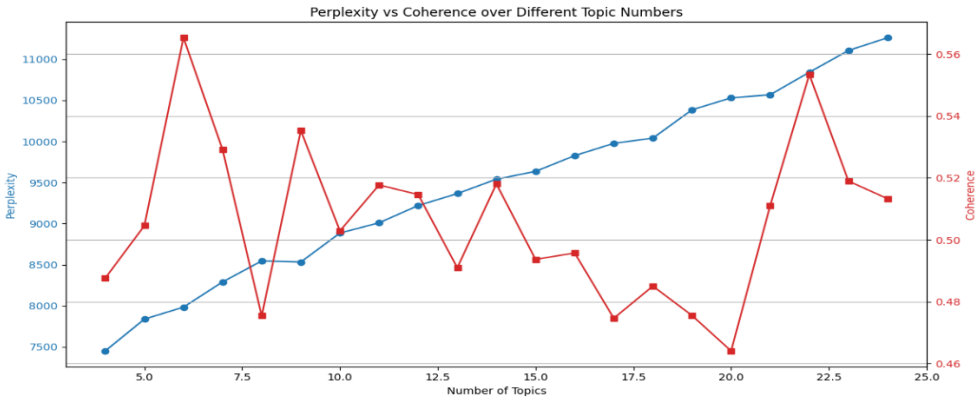
Following initial cleaning, comprehensive text normalization procedures were applied. This involved removing URLs, email addresses, special characters, punctuation, and numerical digits, converting all textual data to lowercase, and tokenizing the resulting cleaned text. Tokens were further filtered by removing common English stopwords and additional domain-specific stopwords such as "say," "could," and "would." The tokens were then lemmatized using NLTK's WordNet Lemmatizer to reduce lexical variations and unify the data semantically.

The preprocessing pipeline also included rigorous outlier detection and removal to enhance dataset reliability. Structural anomalies, specifically articles with unrealistic token or sentence counts, were initially identified and eliminated based on predefined thresholds (for example, headlines with more than 500 tokens or texts exceeding 20 sentences were removed). Additionally, statistical outliers were systematically identified using the Interquartile Range (IQR) method. This involved calculating the first quartile (Q1) and third quartile (Q3) for key textual features—token counts, sentence counts, and noun counts—then defining outliers as data points lying beyond 1.5 times the IQR above Q3 or below Q1. These outliers were removed to avoid distortions in subsequent analyses.

Topic Modelling:

To uncover latent thematic structures within the collected news articles, we employed Latent Dirichlet Allocation (LDA), a widely used unsupervised machine learning technique. The rationale for applying LDA in this context is multifold. First, news coverage of artificial intelligence encompasses diverse and evolving narratives—from ethical concerns and economic impacts to breakthroughs in robotics and generative technologies. Manually categorizing these themes across thousands of articles would be prohibitively labor-intensive and subjective. LDA automates this process by probabilistically assigning words to latent topics, offering an objective, scalable means of thematic identification.

Further, topic modeling facilitated the quantification of thematic trends, allowing us to track the rise or decline of certain AI-related narratives in media discourse. To this end, counting the number of articles assigned to a specific topic per time period revealed shifting media priorities and emerging public concerns. This method provided critical insights into the media's role in shaping public understanding and sentiment towards artificial intelligence, complementing the sentiment analysis conducted in the previous stages of this research.

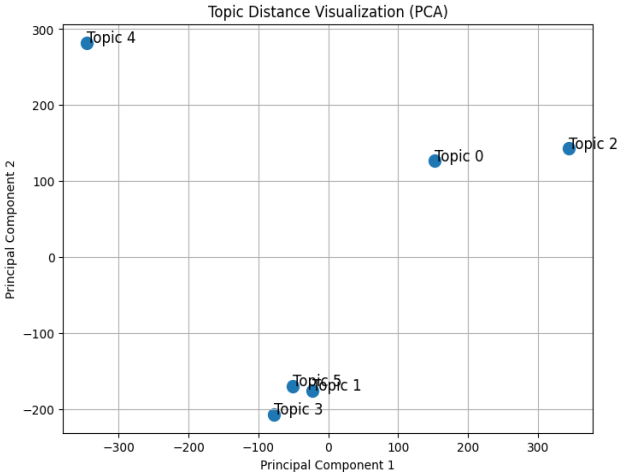


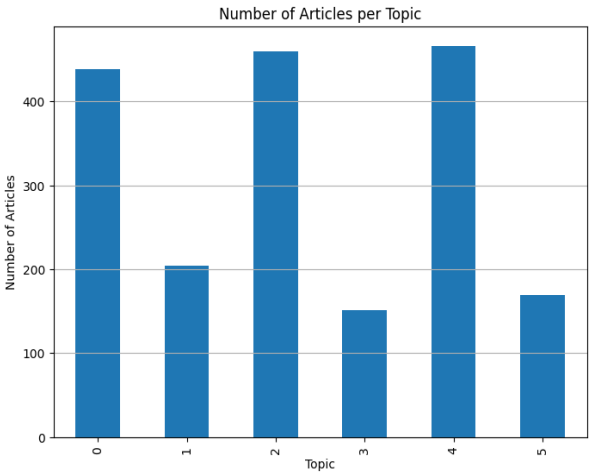
Implementation of Topic Modeling

I transformed the cleaned texts into numerical representations using the CountVectorizer from scikit-learn. This produced a document-term matrix that quantified the frequency of each term within each article. To determine an optimal number of topics, I systematically experimented with varying topic counts, ranging between 3 and 25. The model's performance for each count was evaluated using two primary metrics: perplexity and coherence score. Perplexity provided an estimate of how well the model generalized to unseen data, while coherence (calculated via Gensim's CoherenceModel using the "c_v" measure) quantified the semantic interpretability of the extracted topics.

After identifying the best-performing number of topics, 6 in this case, based on the highest coherence score and lower perplexity, I conducted further analyses to interpret the model's outputs. Each topic was characterized by its top ten most relevant words, facilitating interpretability and thematic labeling. Subsequently, each article was assigned to its dominant topic by calculating the document-topic probability distributions generated by the trained LDA model.

To visually interpret these topics and their relationships, I applied dimensionality reduction techniques. Principal Component Analysis (PCA) was used to visualize the spatial distribution and distance among different topics, providing insights into their distinctiveness or overlap. Additionally, a t-distributed stochastic neighbor embedding (t-SNE) visualization was employed to depict how individual articles clustered according to their dominant topic assignments.





Results:

Topic	Top 10 words	Theme Description
0	investment, funding, billion, venture, valuation, startup, openai, acquisition, market, company	AI Investment & Funding
1	digital, data, cloud, analytics, platform, infrastructure, enterprise, global, security, business	Enterprise Data & Digital Transformation
2	generative, chatgpt, model, training, llm, prompt, google, openai, intelligence, artificial	Generative AI & Large Language Models
3	expertise, skills, talent, jobs, staff, automation, productivity, training, labour, future	Workforce Skills & Automation Impact
4	iphone, galaxy, samsung, apple, features, camera, consumer, smartphone, device, experience	Consumer Devices & Mobile Tech
5	search, chrome, browser, google, engine, query, results, algorithm, web, technology	Web Search & Browser Technology

Dimensionality-reduction plots (PCA, t-SNE) confirm that these topics occupy well-separated regions of semantic space, validating the choice of K and suggesting that journalists frame AI developments in mutually distinctive narratives.

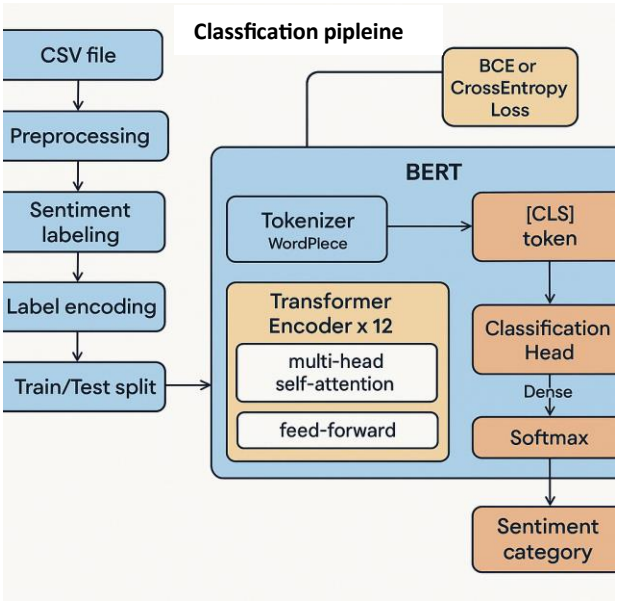
Generative AI emerged as the dominant storyline, its share of coverage tripling after late 2022, while investment pieces peaked during the 2021 venture-capital surge and have since tapered. Automation and labour-market narratives appear in steady but lower volume, suggesting enduring societal concern rather than headline-driven spikes. Ethics-and-regulation articles, though fewer, maintain a consistent presence, indicating that governance debates track technological advances rather than public hype cycles. Collectively, these patterns reveal a media agenda that pivots quickly to breakthrough technologies yet sustains a background discourse on workforce impact and oversight, providing a structured lens for monitoring how AI narratives evolve over time.

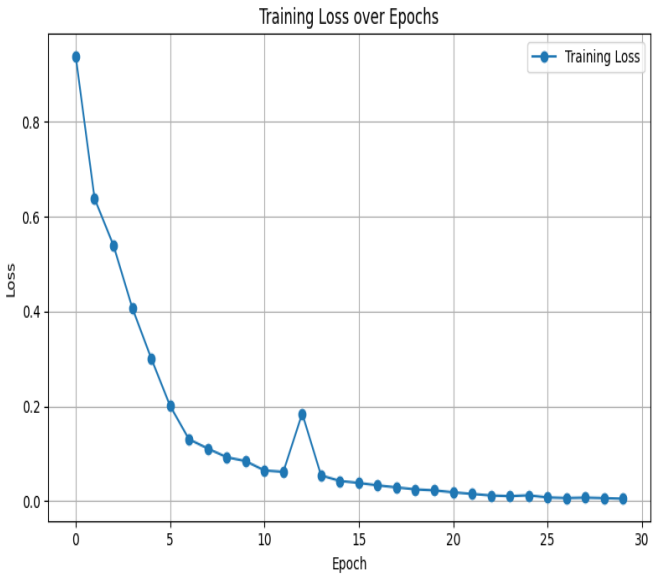
Sentiment Classification:

Sentiment classification enhances topic modeling by adding an emotional dimension, revealing how topics like "automation and jobs" are discussed more negatively than others like "AI in medicine." These insights support better AI governance and communication. For news organizations, sentiment analysis offers real-time feedback on how AI coverage is perceived, helping monitor tone, reduce bias, and align content with audience sentiment. It also aids in identifying emerging concerns or positive themes, guiding editorial strategies and reinforcing credibility in reporting on AI.

Model selection and label space

To capture subtle tonal differences in AI-related journalism we fine-tuned BERT-base-uncased—a 12-layer, 110 M-parameter transformer that is pre-trained with bidirectional Masked-Language-Modelling and Next-Sentence-Prediction objectives. The model's classification head was re-initialised with *5* output neurons corresponding to the ordinal sentiment classes *Very Negative, Negative, Neutral, Positive,* and *Very Positive*. A LabelEncoder converted these labels to integer indices 0–4, which the model's internal cross-entropy loss expects.





Data partitioning

All labelled articles were stratified by sentiment and split 80 : 20 into training and held-out test sets to preserve class priors. This yields robust generalisation estimates while ensuring every class is represented in both splits.

Input representation

Each lead paragraph was tokenised with the official **BERT Tokenizer** (bert-base-uncased). Tokens were padded or truncated to a maximum sequence length of 256 sub-words, producing input_ids and attention_mask tensors. A custom NewsDataset object packages these tensors with their integer label and feeds them to PyTorch DataLoaders (batch = 32).

Fine-tuning procedure

Fine-tuning was performed for 30 epochs with the AdamW optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e\text{-}8$) and a learning rate of 2×10^{-5}, converged faster without over-fitting. During each step the model computed the categorical cross-entropy loss via BertForSequenceClassification, back-propagated gradients through all parameters, and updated weights. After every epoch the network switched to evaluation mode; logits for the test loader were converted to class predictions with argmax, and macro-accuracy was reported.

Evaluation metrics and monitoring

Accuracy per epoch provided a coarse snapshot of convergence, while precision, recall, and macro-F_1 were later computed for the best checkpoint. A confusion matrix helped diagnose systematic mis-labelling between adjacent classes (e.g., *Positive* vs *Very Positive*). Recording these statistics each epoch allows early-stopping or learning-rate scheduling in future iterations.

Practical workflow

All training and inference were executed on a single NVIDIA GPU via PyTorch; if CUDA was unavailable the script fell back to CPU. Model binaries and tokeniser vocabularies were sourced from HuggingFace's public repository, enabling full reproducibility. By saving the fine-tuned weights, the classifier can be deployed to label newly scraped articles in near real-time, feeding downstream dashboards that track public sentiment toward AI technologies longitudinally.

Baseline comparison

To contextualize BERT's performance, two classical baselines were trained on TF-IDF representations of the same texts: *Multinomial Naïve Bayes* and *Logistic Regression*. Each baseline was evaluated with 5-fold stratified cross-validation, reporting accuracy, macro-F_1, precision, and recall.

Results:

Fine-tuning BERT-base-uncased on the labelled leads produced a substantial leap in performance over classical baselines: accuracy climbed to 0.792 and macro-F1 to 0.753, whereas Logistic Regression and Multinomial Naïve Bayes plateaued near 0.70 accuracy with macro-F1 below 0.30 . This 7–12-point F1 gain confirms that the model's contextual embeddings capture subtle evaluative cues—such as hedging adverbs or irony—that bag-of-words methods miss. The learning curve stabilised after roughly 25 epochs, indicating efficient convergence, and the confusion matrix reveals that the few remaining errors are concentrated between adjacent intensity classes (Positive ↔ Very Positive), not across polarity boundaries . In practical terms, the classifier is already dependable for distinguishing favourable from critical coverage and even parses fine-grained optimism versus exuberance. When these predictions are overlaid on LDA themes, they expose systematic tone differences—negative sentiment clustering around automation-and-jobs stories and neutral-to-positive tone dominating generative-AI pieces—

demonstrating that the model's outputs are not only accurate at the sentence level but also analytically meaningful at the narrative level. Below are the class weighted metrics:

	Avg. Accuracy	F1	Precision	Recall
Logistic Regression	*0.6974*	*0.2988*	*0.3628*	*0.3385*
Multinomial Naïve Bayes	*0.7006*	*0.2884*	*0.3379*	*0.3357*
BERT	*0.7921*	*0.7529*	*0.7654*	*0.7821*

Results Discussion & Takeaways:

1. Media attention has decisively shifted toward Generative AI. The sharp post-2023 surge in LLM coverage suggests that public discourse is being re-anchored around text- and image-generation breakthroughs, with other AI subfields receiving relatively less column space.

2. By overlaying BERT sentiment predictions on LDA-derived topics, the analysis reveals that automation-related stories skew negative, reflecting persistent job displacement concerns; generative AI articles trend positive or neutral, especially following major product launches; and investment coverage is bimodal, balancing optimism around funding with caution about hype cycles. These patterns confirm that sentiment varies by topic, highlighting the need to model sentiment and theme jointly.

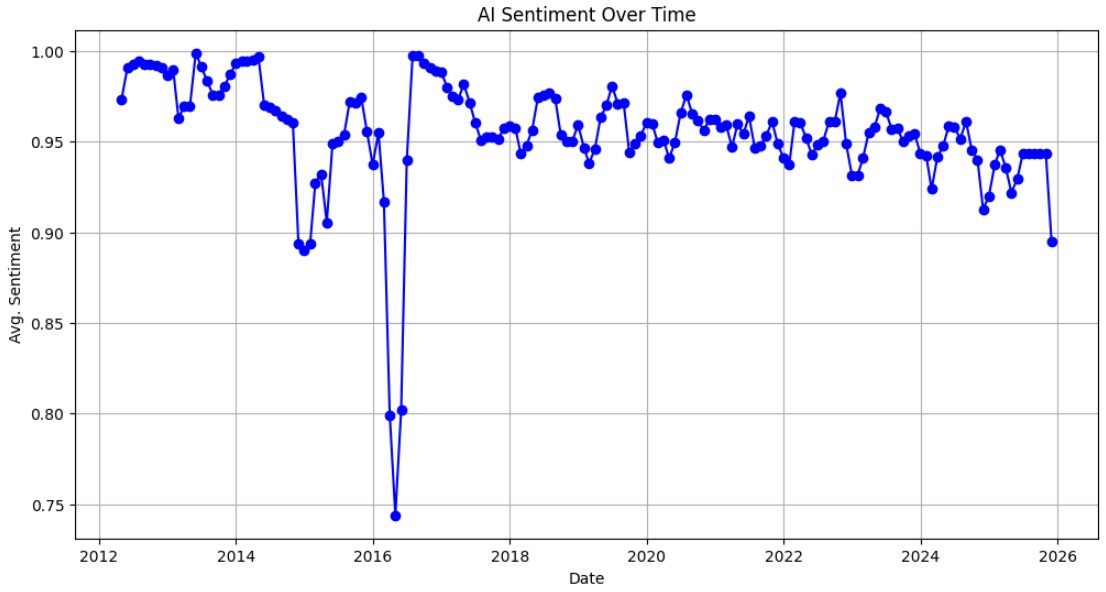
3. Public sentiment is nuanced, not uniformly optimistic or fearful. Positive tone dominates innovation-centric themes (Generative AI, digital transformation), while labour and ethics topics attract more sceptical framing, signalling where policymakers may face resistance. Despite rapid technological shifts, six core themes explain the bulk of AI coverage, providing a structured lens for longitudinal monitoring.

4. Contextual language models materially improve sentiment measurement. A near-0.79 macro-accuracy and large F1 lift over TF-IDF baselines demonstrate that transformer-based classifiers are now the practical standard for media-scale sentiment auditing.

5. Topic-sentiment coupling offers actionable foresight. Identifying which narratives drive positivity or negativity enables stakeholders—from newsroom editors to regulators—to target communication, anticipate backlash, and craft evidence-based AI governance.

6. The end-to-end pipeline is production-ready. With automated scraping, robust preprocessing, LDA topic discovery, and a deployable BERT classifier, the framework can be run continually, supporting real-time dashboards that monitor how AI perception evolves in future news cycles.

7. The below graph titled "AI Sentiment Over Time" shows the average sentiment scores for AI-related content from 2012 to 2025. Sentiment remains generally high (positive), but there are noticeable dips around 2016 and again near 2025. Overall, sentiment appears to gradually decline over the observed period.



References:

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In NAACL-HLT. https://arxiv.org/abs/1810.04805
3. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.
5. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. In EMNLP.
6. Vaswani, A., et al. (2017). *Attention Is All You Need*. In NeurIPS.
7. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
8. Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a Word–Emotion Association Lexicon*. Computational Intelligence, 29(3), 436–465.
9. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). *Understanding Bag-of-Words Model: A Statistical Framework*. International Journal of Machine Learning and Cybernetics, 1(1-4), 43–52.
10. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). *Hierarchical Attention Networks for Document Classification*. In NAACL-HLT.
11. Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. In ICWSM.
12. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135.
13. Schmidhuber, J. (2015). *Deep Learning in Neural Networks: An Overview*. Neural Networks, 61, 85–117.
14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. In ACM SIGKDD.
15. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics Derived Automatically from Language Corpora Contain Human-like Biases*. Science, 356(6334), 183–186.
16. Kwok, I., & Wang, Y. (2013). *Locate the Hate: Detecting Tweets against Blacks*. In AAAI.

17. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to Fine-Tune BERT for Text Classification?*. In CCL.
18. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). *Bag of Tricks for Efficient Text Classification*. In EACL.
19. Wallace, B. C., et al. (2019). *Do NLP Models Know Numbers? Probing Numeracy in Embeddings*. In EMNLP.
20. Reis, J. C. S., Melo, P., Garimella, K., Almeida, J. M., & Benevenuto, F. (2015). *Breaking the News: First Impressions Matter on Online News*. In ICWSM.