

Propaganda and Fake News Detection using Machine Learning and Deep Learning from News and Social Network Corpora



STUDENT NAME: MUHAMMAD HAROON
STUDENT ID: 2434025
SUPERVISOR: MUHAMMAD AZEEM ABBAS

Table of Contents

..... 0

Abstract..... 2

1. Introduction..... 3

2. Literature Review 4

 2.1. Traditional Machine Learning Approaches..... 4

 2.2. Deep Learning and Transformer-Based Architectures 4

 2.3. The Challenge of Cross-Domain Generalization 5

 2.4. Real-World Impact: The Southport Case 5

 2.5. Research Gaps..... 5

3. Objectives 7

4. Importance of Research..... 7

5. Research Questions and Hypothesis 7

6. Research Design and Methodology 8

 6.1. Phase 1: Data Acquisition & Preprocessing 8

 Preprocessing Pipeline..... 8

 6.2. Phase 2: Feature Extraction & Model Implementation 8

 6.3. Phase 3: Model Training and Evaluation..... 8

 6.4. Phase 4: Cross-Domain Generalization 8

 6.5. Phase 5: Analysis and Thesis Writing 9

7. Research Limitations 9

8. Timeline with Professional Practice..... 9

9. Expected Findings 10

References..... 11

Abstract

In this study, I investigate the capabilities of machine learning and deep learning algorithms to identify propaganda and fake news in text-based online information. It is a response to the problem of digital disinformation which can have a significant real-world impact as was evident during the Southport riots in 2024. The three traditional machine learning models (Support Vector Machine, Logistic Regression, Multinomial Naïve Bayes) will be compared to deep learning models (LSTM model and lightweight BERT model). The models will be trained and tested on the following benchmark datasets: LIAR (12.8k political statements), WELFake (72k news articles) and a Social Media Propaganda dataset (10k-15k tweets).

The accuracy, precision, recall, F1-Score, and inference speed will be used to measure the performance. The research will also explore the generalizability of the models by training on one set of data and testing the models on the other. We will be using Google Colab for all the experiments.

The anticipated result will be a clear comparison of the models, with the one that is most appropriate for real-time fake news and propaganda detection being identified.

1. Introduction

Social networks are now one of the quickest means of spreading information in the digital era. But the same tools that are available for communication and information are also being utilized for the dissemination of disinformation and manipulation of public opinion (Aïmeur, Amri and Brassard, 2023). Competing groups are seeking to manipulate storylines in social media, and at times, truth is sometimes overlooked in the process. The threat of online propaganda is no longer a theoretical issue; it is an actual problem affecting both people and technology that needs to be addressed and solved. It can propagate rapidly and widely in the online world, and particularly rapidly on social media platforms, thanks to algorithmic amplification which can spread to millions of people in hours (Huszár *et al.*, 2022).

One recent instance of this was the Southport stabbing in July 2024. A deleted LinkedIn post suggested the migrant was to blame and that claim was viewed over 3 million times on X – formerly known as Twitter. The research, conducted by the BBC, revealed that this misinformation was connected to riots in England and Northern Ireland, and that propaganda is not just false but can also cause real-world harm, as (Thomas and Sardarizadeh, 2024) prove. The false claim was typical, as it was first posted, then frozen for screenshots, amplified by verified accounts, including far-right ones like Paul Golding and Laurence Fox, and then presented as credible evidence by some users despite being debunked. One key finding of this case is the fact that misinformation can readily outpace efforts from platforms to moderate content and any official corrections.

The aim of this research is to develop automated systems that can identify harmful misinformation in a timely and accurate manner. Social media moderation techniques are unable to keep up with viral false claims, particularly when they're being spread by influencers or accounts that have paid verification badges. In the case of Southport, the name “Ali-Al-Shakati” was also common knowledge before Merseyside Police confirmed the man was a 17-year-old Cardiff-born man. The gap in the dissemination of misinformation and the dissemination of official information is where automated detection tools could come in handy (Thomas and Sardarizadeh, 2024). So computational means of identifying propaganda and fake news are becoming more and more crucial. This thesis will be a 3-month project with a practical and reproducible framework which can be implemented using available tools like Google Colab.

Research on this field has progressed but there are drawbacks. Large propaganda detection datasets often rely on weak labels, in which the content is tagged according to its source, not by a careful check of its contents. For instance, a post could be classified as propaganda only because it is from a publisher that is associated with propaganda. Weak labels share only about 41% of the same examples with human labels (Maarouf *et al.*, 2024a), suggesting that there are many examples that are labeled weakly which may be unreliable. Their study also reveals that models trained on weak labels only, yield an AUC of just 64.03%, while the models obtained by the high-quality human annotations achieve an AUC of 92.25%. It indicates that the architecture of the model is not necessarily the reason for the poor performance, but may be due to training on data that is noisy and has poor reliability.

Another challenge in the literature is that numerous studies concentrate on a specific social media platform or just news articles (Bisiani and Heravi, 2023). But misinformation can cross platforms in the real world. The false claim started on LinkedIn, got shared with screenshots on X and was then discussed on Telegram and other communication platforms in the Southport case. This demonstrates the importance of more generalizable detection systems in the different domains and types of content. In order to fill this gap, this study is going to implement a cross-domain generalization experiment with three datasets.

The study will compare five different models in detail on 3 representative datasets. Traditional machine-learning classifiers (Support Vector Machine, Logistic Regression, Multinomial Naïve Bayes) and deep learning models (LSTM) are included in these models. The outcome is expected to be a practical comparison of the approaches, focusing especially on the challenges they face in real-world scenarios, including speed, accuracy, and cross-domain performance when it comes to detecting propaganda and fake news.

2. Literature Review

The study of digital disinformation has advanced from an understanding of how fake information spreads on the internet to an automated approach to deciding how to spot it. This literature review consolidates research from three key streams: traditional machine learning techniques, deep learning and transformer-based techniques, and the persistent issue of cross-domain generalization.

2.1. Traditional Machine Learning Approaches

The early literature on identification of propaganda or fake news heavily used traditional machine learning methods and manually engineered features. Based on the work of Zhang (2025), who introduced the classical classifiers as a foundation for detecting propaganda on social media by applying TF-IDF feature extraction, (Zhang, 2025) used TF-IDF feature extraction and classical classifiers to detect propaganda on social media. Their study revealed that SVM gave the best performance with an accuracy of 69.84%, followed by Multinomial Naïve Bayes with accuracy of 69.27% and Logistic Regression with accuracy of 68.76%. The experiments were performed on a self-annotated set of about 5000 tweets, and features like unigrams, bigrams, and TF-IDF representations were used.

More recently, wider surveys of the field have again reiterated the continued importance of machine learning, but also the benefits of integrating machine learning with deep learning approaches. A systematic review conducted by (Plikynas, Rizgelienė and Korvel, 2025) revealed that hybrid ML and DL methods hold promise for detecting fake news, propaganda, and disinformation on various dimensions, such as content properties, propagators, and social effects. Although this progress has been made, there remain significant drawbacks to traditional methods. Often requires a lot of feature engineering; sensitive to class imbalance; often has poor transfer learning ability from one domain to the other.

2.2. Deep Learning and Transformer-Based Architectures

Over the past few years, the field of detecting misinformation has undergone a transformation thanks to the introduction of deep learning, including transformer-based architectures. These models are capable of learning the context for a specific term more effectively than traditional models, and have performed very well in many applications of NLP.

A significant contribution by (Maarouf *et al.*, 2024), who created the first large-scale, human-annotated dataset for the detection of online propaganda, which has around 30,000 samples. Their results revealed that quality of data is a significant factor affecting the performance of the model. Weak labels only matched human level annotations about 41% of the time, but models trained on high quality labels performed quite well. BERTweet fine-tuned on human-annotated data outperforms the other models with an AUC of 92.25% in their experiments.

Future studies have focused on hybrid architectures using sequential learning methods and transformer models. It was shown that the use of BERT in conjunction with LSTM networks can enhance the ability to

detect misinformation by leveraging the network's contextual understanding of language with its sequence modeling capabilities. (Keya *et al.*, 2023) introduced a hierarchical network model called FakeStack, which integrates the parts of Tri-BERT, CNN and LSTM. They achieved some excellent results on various datasets such as 75.58% accuracy on LIAR and 98.25% accuracy on WELfake, making hybrid deep learning approaches promising.

2.3. The Challenge of Cross-Domain Generalization

Though many studies have shown high performance, an important question is: How well do models generalize in un-training environments?

To address this, (Keya *et al.*, 2023) assessed the performance of five deep learning models on various datasets. They found that there was a large difference between within domain and across domain performance. Models performed well with data that was comparable to their training distribution, with an accuracy of close to 99%, but upon testing with different data set, the accuracy drops by about 30%. The results demonstrate the challenge of developing systems that are reliable enough to work on various types of content and platforms.

(Alnabhan and Branco, 2024) suggested BERTGuard, a multi-stage system that uses the knowledge of what the content is before employing domain-specific BERT models. This approach boosted the robustness of the model in various aspects. A few other recent works have looked into hybrid transformer models that integrate models like DeBERTa, T5, and BART with BiLSTM attention mechanisms, and found that they achieve better performance on various Twitter-based misinformation datasets.

2.4. Real-World Impact: The Southport Case

The relevance of this research area is seen in the real-life situations. The BBC report on the 2024 Southport riots (Thomas and Sardarizadeh, 2024) detailed the manner in which a false claim, which first appeared in a LinkedIn post, rapidly proliferated in social media platforms and was shared by millions of users before official details were released. The incident showed the power of misinformation to impact public opinion and the shortcomings of existing moderation frameworks. It also highlighted the importance of automated detection methods which are able to work on multiple platforms and quickly.

2.5. Research Gaps

While existing literature has brought in effective detection models, there are some important gaps.

First, there are significant differences in the way the items have been evaluated in the different studies, so direct comparison is difficult. Second, cross-domain generalization is under explored, especially by conducting systematic experiments on performance deterioration across datasets. Third, although response time is important for real-time deployment, relatively few studies report inferences speed.

Table 1: Research Gaps

Approach	Arithmetic Algorithm	Algebraic Algorithm	Classifier Models	Datasets	Accuracy	Remarks
Traditional Machine Learning (Hasan, Roy and Nafis, 2025)	TF-IDF, Bag-of-Words (BoW)	Linear / Probabilistic	SVM, Logistic Regression, Random Forest, Extra Trees, XGBoost, Naive Bayes	LIAR (political statements)	Binary: 0.624 (SVM) Multi-class: 0.262 (Extra Trees)	Demonstrates a "performance ceiling" for text-only models on political fact-checking. Linear SVM matches RoBERTa, proving model capacity is not the bottleneck
Deep Learning (LSTM/CNN) (Machová, Mach and Balara, 2025)	Word Embeddings (GloVe)	Sequential / Convolutional	LSTM, BiLSTM, CNN, GRU, CLSTM	LIAR (12.8k statements)	Max Accuracy: ~0.27 (6-class)	Attention mechanisms improve accuracy on LIAR only marginally due to dataset limitations (short texts, semantic ambiguity).
Deep Learning (LSTM/CNN) (Borse, Kharate and Kharate, 2025)	Word Embeddings (GloVe, fastText)	Attention-based Seq2Seq	CNN, LSTM, BiLSTM, HAN, GAN, GCN	LIAR, ISOT, WELFake, FakeNews Net	Varies by dataset	Comprehensive survey of DL architectures for fake news verification. Highlights strengths/weaknesses of each model type
Transformer Models (Kuntur et al., 2024)	Subword Tokenization (BPE), Positional Encoding	Self-Attention, Multi-Head Attention	BERT, RoBERTa, GPT-2	FakeNewsNet, ISOT, WELFake	FakeNewsNet: 86.16% (RoBERTa) ISOT: 99.99% (RoBERTa) WELFake: 99.72% (GPT-2)	Transformers significantly outperform GNNs on text classification. BERT/RoBERTa achieve near-perfect accuracy on large, balanced datasets like ISOT and WELFake.
Large-Scale Benchmarking (Dell'Oglio et al., 2026)	TF-IDF, Sentence-BERT	Multi-Domain Evaluation	12 models (ML, DL, Transformers, LLMs, Cross-Domain)	10 datasets (news, social media, political statements)	In-domain: High (e.g., 99%+) Cross-domain: Significant drop (20%+)	Critical finding: Models overfit to dataset-specific artifacts. Specialized cross-domain architectures generalize better but are data-hungry; LLMs offer

To address these limitations, five models are implemented and tested under the same experimental conditions, the performance is compared on three different datasets, and the performance change along the domains is measured, besides the usual classification performance measures, the inference speed. This will help to offer a fairer and more realistic understanding of propaganda and fake news detection techniques.

3. Objectives

Table 2: Research Objectives

Objective	Description
Objective 1 (Implementation)	Design and benchmark three classical machine learning models such as Support Vector Machine, Logistic Regression and Multinomial Naïve Bayes, and deep learning model LSTM, to detect fake news and propaganda in text.
Objective 2 (Evaluation)	The performance of each model is evaluated using commonly used categorization metrics such as accuracy, precision, recall and f1 score, and is also compared on inference time per sample to determine which model is better for real-time detection.
Objective 3 (Generalization)	To analyze the performance of the models on the three different datasets (LIAR, WELFake, a Social Media Propaganda containing fake news) to see how well they generalize across the text domains.
Objective 4 (Reproducibility)	To generate a fully documented Google Colab codebase that enables the experiments to be replicated, reviewed and extended in future research. Objective 4: Reproducibility

4. Importance of Research

This is a scientific and practical research. As a scientific approach, it will analyze and contrast traditional machine learning models with the modern deep learning models for detecting propaganda and fake news with a focus on cross-domain generalization (Alnabhan, 2025). The study will test the models on three types of data, political news statements and social media posts, to better demonstrate their ability to be applied beyond the data they were trained on. It will also analyze the trade-off between accuracy and inference speed that is critical in designing real-time detection systems.

The results have implications in practice, such as for social media companies, fact-checkers and policy makers to select models better suited for reliable detection and timely response. It is particularly important where misinformation is rapidly disseminated and has the potential to result in harm in the real world, like the riots in Southport. The project will also result in a fully documented implementation in Google's Colab environment that will facilitate the reproduction, adaptation and extension of the work by other researchers and practitioners. The three months allow for a thorough and useful contribution, but not too long for a study topic.

5. Research Questions and Hypothesis

Table 3: Research Questions and Hypothesis

Research Question	Hypothesis
RQ1: What are the comparative performances of traditional ML models (SVM, LR, MNB) and a deep	H1: The deep learning model (LSTM) will achieve the highest overall F1-score (estimated 0.78–0.82),

learning model (LSTM) in detecting propaganda across different text datasets?	outperforming SVM (estimated F1 < 0.75) and other ML models.
RQ2: Which model architecture offers the best balance between detection accuracy and inference speed for real-time deployment in a Google Colab environment?	H2: LSTM will provide the best accuracy-speed balance, while traditional ML models will be faster but less accurate.
RQ3: How well do models generalize when trained on one dataset and tested on another (LIAR, WELFake, Social Media)?	H3: All models will show significant F1-score decline across domains (estimated 15–30% drop). TF-IDF-based ML models will suffer most due to lack of contextual understanding.

6. Research Design and Methodology

A quantitative experimental approach is taken in this study, and all of the work was done with the help of Python and Google Colab.

6.1. Phase 1: Data Acquisition & Preprocessing

Table 4: Data Acquisition & Preprocessing

Dataset	Size	Content Type	Label Format	Binary Conversion
LIAR	12.8k	Short political statements	6 truth levels	pants-fire/false → Propaganda (1); others → 0
WELFake	72k	News articles	Real / Fake	Fake → Propaganda (1); Real → 0
Social media	10k-15k	Tweets	Binary (human-annotated)	Direct use as provided

Preprocessing Pipeline

This pre-processing stage will remove any URLs, mentions of people like @user and special characters. Consistent, all text will be converted to lower case. For traditional machine learning models, stopwords will be stripped out, but for deep learning model they will be kept so that the models could account for sentence structure and context. Class imbalance will be addressed in the training of the models with class weights, and SMOTE will be used in machine learning models as needed.

6.2. Phase 2: Feature Extraction & Model Implementation

Table 5: Feature Extraction and Model Implementation

Model	Feature / Architecture	Implementation
SVM, LR, MNB	TF-IDF (unigram + bigram, max_features=10000)	Scikit-learn
LSTM	Embedding (128) → LSTM (64, return_sequences=True) → Dropout (0.5) → LSTM (32) → Dense (1, sigmoid)	TensorFlow/Keras

6.3. Phase 3: Model Training and Evaluation

- As for Train/Test splitting, it was done based on the stratified train split within each random dataset independently with 70/30 ratio.
- Two sets of metrics shall be covered: Accuracy, Precision, Recall, F1-Score, Confusion Matrix
- Objective: Inference speed: ms sample (Average of 1000 samples)

6.4. Phase 4: Cross-Domain Generalization

- All five models are first tested to see if they can be transferred to the dataset sets LIAR, WELFake and social media, to test generalization.

- The performance measures used to measure the models are Accuracy, Precision, F1 Score, Support indicated in the matrix.

6.5. Phase 5: Analysis and Thesis Writing

- Performance of models will be compared based on statistics significant tests employing paired t-tests at 5% level.
- The results will be presented in comparison tables and pictograms such as bar charts and heatmaps.
- The thesis will be developed throughout the project in parallel with experimental works to achieve a steady thesis development all along.

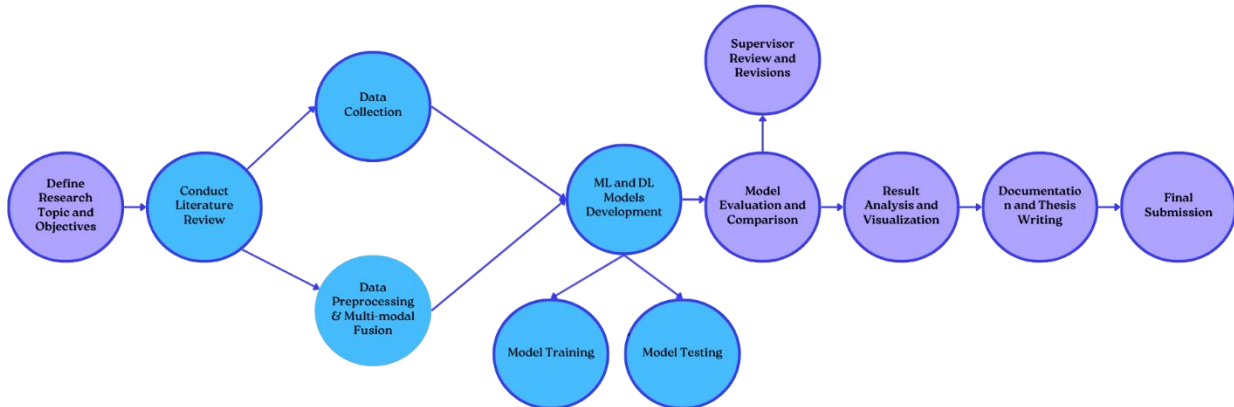


Figure 1: Workflow Diagram

7. Research Limitations

The data utilized in this study have important limitations and are varied. They are not entirely representative of real-world propaganda, which may consist of images, videos, memes and deepscans. Given the fact that the LIAR data is a dataset of short political statements, on average 17 words in length, it may not represent the complexity of longer propaganda narratives. WELFake is large with a major focus on English-language Western news, so this may be of less value in other English or non-English, and non-Western contexts. The Specificities of each social media platform (e.g. Twitter's character limit) could also potentially lead to data being unrepresentative of social media sites like Facebook, TikTok and Instagram.

In terms of the methodology, a constraint of the study is the period of observation, which was 3 months. Therefore, the hyperparameter tuning will be based on some reasonable default values instead of sweeping all the possibilities.

Generalizability limitations also exist with the study. These models are fed with historical data including LIAR, WELFake and social media data prior to 2023 so they may not be able to identify more recent propaganda methods or happenings. For evaluation, the real-time detection will be performed on static data but not on live on airflow.

8. Timeline with Professional Practice

Project Period: June 1st – August 28th (12 weeks)

Table 6: Project Timeline

Week	Dates	Task
------	-------	------

01	June 1-7	Literature review finalization; GitHub setup
02	June 8-14	Download all 3 datasets; preprocessing pipeline
03	June 15-21	TF-IDF + ML models (SVM, LR, MNB)
04	June 22—July 5	LSTM model implementation
05	July 6-12	Within-domain evaluation (all 3 datasets)
06	July 13-19	Cross-domain experiments (6 transfers)
07	July 20-26	Inference speed benchmarking
08	July 27-Aug 2	Draft Results and Discussion
09	Aug 3-9	Draft Introduction, Methodology, Conclusion
10	Aug 10-16	Revise entire thesis; create final figures
11	Aug 17-28	Final proofreading; submission



Figure 2: Gantt Chart

9. Expected Findings

The overall performance of Deep Learning is likely to be best on the tasks within its language domain. It is expected to work best on WELFake, since there are many more samples as well as longer texts with the type of articles. I expect it to be about as good on the Social Media data set as well, albeit slightly worse due to the nature of tweets being very casual and short, and consisting of conversation. The effect of low-content statements in LIAR can be greater than in other tasks as the statements vary significantly in their length, and typically have an average of only about 17 words.

LSTM is expected to have moderate performance across all three datasets, especially on WELFake, and also on social media. It may perform poorly on LIAR though, as there is not much of a context for short statements.

Traditional machine learning algorithms, such as SVM, Logistic Regression and Multinomial Naïve Bayes, are likely to perform on the bottom of the pile. Although SVM is the best traditional model, this is still anticipated to lag behind LSTM, particularly on LIAR.

With cross domain generalization all six transfer tasks should significantly decrease their F1 score. The biggest drops are likely to be observed when models are tested on WELFake or social media, as the short-length political claim may not capture the richer patterns in the news or social media text. Tighter falls are anticipated going from WELFake over social media. The resulting models are expected to perform best for LSTM, while the performance of traditional models is expected to be the worst based on the number of features they have that are specific to the given dataset.

LSTM should have the most accurate inferences and be the fastest in real-time suitability. In some time-sensitive applications, the resource LSTM may not be as accurate but might be quicker. Models are assumed to be the most efficient, but less accurate, candidates for practical use.

References

Aïmeur, E., Amri, S. and Brassard, G. (2023) "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, 13(1), p. 30. Available at: <https://doi.org/10.1007/s13278-023-01028-5>.

Alnabhan, M.Q. and Branco, P. (2024) "Bertguard: two-tiered multi-domain fake news detection with class imbalance mitigation," *Big Data and Cognitive Computing*, 8(8), p. 93.

Alnabhan, M.Q.M. (2025) *Advancing Cross-Domain Fake News Detection: Enhanced Models to Improve Generalization and Tackle the Class Imbalance Problem*. PhD Thesis. Université d'Ottawa/University of Ottawa. Available at: <https://ruor.uottawa.ca/server/api/core/bitstreams/2d071148-bfe8-40be-88c1-c9be567ef558/content> (Accessed: May 22, 2026).

Bisiani, S. and Heravi, B. (2023) "Uncovering the state of local news databases in the UK: Limitations and impacts on research," *Journalism and Media*, 4(4), pp. 1211–1231.

Borse, A.A., Kharate, G.K. and Kharate, N.G. (2025) "Deep Learning-Powered Approaches for News Authenticity Verification and Fake News Detection: A Comprehensive Survey," *Advances in Nonlinear Variational Inequalities*, 28(7).

Dell'Oglio, P. *et al.* (2026) "An Experimental Comparison of the Most Popular Approaches to Fake News Detection." arXiv. Available at: <https://doi.org/10.48550/ARXIV.2603.25501>.

Hasan, S.M., Roy, S. and Nafis, A.J. (2025) "Generalization Gaps in Political Fake News Detection: An Empirical Study on the LIAR Dataset." arXiv. Available at: <https://doi.org/10.48550/ARXIV.2512.18533>.

Huszár, F. *et al.* (2022) "Algorithmic amplification of politics on Twitter," *Proceedings of the National Academy of Sciences*, 119(1), p. e2025334119. Available at: <https://doi.org/10.1073/pnas.2025334119>.

Keya, A.J. *et al.* (2023) "FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection," *Plos one*, 18(12), p. e0294701.

Kuntur, S. *et al.* (2024) "Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study," *Electronics*, 13(23), p. 4784. Available at: <https://doi.org/10.3390/electronics13234784>.

Maarouf, A. *et al.* (2024a) "HQP: A Human-Annotated Dataset for Detecting Online Propaganda," *Findings of the Association for Computational Linguistics ACL 2024*, pp. 6064–6089. Available at: <https://doi.org/10.18653/v1/2024.findings-acl.363>.

Maarouf, A. *et al.* (2024b) "HQP: a human-annotated dataset for detecting online propaganda," *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6064–6089. Available at: <https://aclanthology.org/2024.findings-acl.363/> (Accessed: May 22, 2026).

Machová, K., Mach, M. and Balara, V. (2025) "Analysis of the Effect of Attention Mechanism on the Accuracy of Deep Learning Models for Fake News Detection," *Big Data and Cognitive Computing*, 9(9), p. 230. Available at: <https://doi.org/10.3390/bdcc9090230>.

Plikynas, D., Rizgeliene, I. and Korvel, G. (2025) "Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning," *IEEE Access*, 13, pp. 17583–17629.

Thomas, E. and Sardarizadeh, S. (2024) *Southport riot: How a LinkedIn post helped spark unrest - BBC tracks its spread, How a deleted LinkedIn post was weaponised and seen by millions before the Southport riot*. Available at: <https://www.bbc.com/news/articles/c99v90813j5o> (Accessed: May 22, 2026).

Zhang, L. (2025) "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification," *PLoS One*, 20(7), p. e0327347.