IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

News Classification at Scale: A Web Scraping and BERT-Based Approach

Ms. Khushali Domadiya
Lecturer, CE-IT, SOE, P P Savani University, Surat, India

Abstract: *The quick growth of digital news media resulted in the accumulation of a tremendous amount of unstructured text data. The manual collection and classification of such data is not only inefficient but also hard to scale. The current research introduces a fully automated pipeline which combines web scraping with machine learning and deep learning techniques for large-scale news classification. The system encompasses automatic scraping, preprocessing, feature engineering, and classification through both traditional models and transformer-based architectures. The experiments conducted on the dataset of more than 20,000 articles show that the combination of automated scraping and the use of advanced NLP models like BERT has markedly improved the accuracy of classification as well as the efficiency of the whole operation. Furthermore, the paper presents the system architecture, experimental evaluation, challenges, and suggestions for future improvements.*

Keywords: NLP, Web Scraping, BERT, Beautifulsoup4, selenium, SVM, Logistic Regression, Random Forest, LSTM

I. INTRODUCTION

Digital journalism has evolved in such a way that now there are millions of articles produced every day through portals, blogs, and online media outlets. It is true that the vast amount of information supports to keep the news updated but at the same time brings in difficulties thinking of organizing and retrieving the proper content. The mentioned groups, which are researchers, media organizations, and automated systems, need the news in structure and categorized forms for the purpose of analysis, trend prediction, and information retrieval.

With modern scales, manual news collection is impossible to accomplish. The use of automated web scraping allows for the continuous, structured extraction of content, and at the same time machine learning and deep learning provide the classification of the vast amounts of data as a whole. The present study looks into a single pipeline that is able to automate the complete process, from data gathering to classification. The research paper describes the methodology, the experimental results, and the entire success of merging automation with NLP in the context of news categorization on a grand scale.

II. LITERATURE REVIEW

The prior works have carried out vast amounts of research on text classification mainly using traditional models such as Naïve Bayes, Logistic Regression, SVM, and Random Forest, coupled with Bag-of-Words or TF–IDF representations. These kinds of methods yield good results on datasets with moderate size but they fall short in understanding the semantics on a deeper level.

Overfitting reduction in deep learning models such as LSTMs and BiLSTMs, which captured contextual dependencies, was one of the main reasons for their usage in sequence modeling. Yet, transformer-based models, like BERT, RoBERTa, and DistilBERT, brought about a significant change in classification tasks by their attention-based bidirectional contextual comprehension.

The web scraping methods which implement BeautifulSoup, Scrapy, and Selenium provide unstructured web content extraction in a structured way, however, the research done so far usually centers either on scraping or on the classification aspect. There are very few studies that present a fully automated pipeline covering the entire process of extraction,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

cleansing, feature engineering, and classification. This is precisely the shortcoming that is being rectified by the present study.

III. METHODOLOGY

The system that has been suggested is made up of four basic parts:

1. Automated Web Scraping
2. Data Cleaning and Preprocessing
3. Feature Extraction
4. Classification Models

III.I Automated Web Scraping

Automated scraping gathers up every bit of information like:

- Headlines
- Article content
- Timestamps
- Author information
- Tags
- Source links

There were different methods of scraping applied to tackle the various structures of the websites:

- Direct extraction of HTML
- Processing of JavaScript-rendered content
- Covering of paginated areas and crawling
- Setting up regular scraping for getting constant updates

Consideration of ethics included using robots.txt and limiting the rate of requests as per the regulations.

III.II Data Cleaning and Preprocessing

Scraped data was polluted by HTML tags, embedded ads, inconsistent formatting, and duplicated entries. The cleaning process included:

HTML removal
Lowercasing
Removal of unnecessary symbols and numbers
Duplicate elimination
Sentence segmentation
Stopword removal
Lemmatization

Preprocessing that was effective led to the enhancement of feature quality as well as model performance.

III.III Feature Extraction

Three feature extraction methods were applied:

TF-IDF: It was very useful for classic machine learning models.
Word Embeddings: Word2Vec and GloVe demonstrated the semantic relations.
Transformer Embeddings: The contextual embeddings based on BERT granted the best sentence-level representation, which made the deep learning classification better.

III.IV Classification Models

Two model categories underwent an evaluation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

A. Conventional Machine Learning Models
• Naïve Bayes
• Logistic Regression
• SVM
• Random Forest
B. Deep Learning Models
• LSTM
• BiLSTM
• BERT / DistilBERT
The superiority of the transformers over the other models was primarily based on their ability to understand the context.

IV. PROPOSED SYSTEM ARCHITECTURE

The system consists of several components:
1. Web Scraper: It collects and stores data regarding the articles.
2. Preprocessing Engine: This is where the text is cleaned and standardized.
3. Feature Extraction Layer: TF-IDF vectors or embeddings are created here.
4. Classification Engine: The use of ML or DL models takes place.
5. Evaluation Module: Accuracy, precision, recall, and F1-score are calculated.
6. Deployment Layer: For real-time classification, an API-based interface is provided.
The modular design of the system permits easy scaling and integration into real-world applications.

V. EXPERIMENTAL RESULTS

From a set of 20,514 news articles scraped, a total of seven categories were selected: Politics, Sports, Business, Technology, Entertainment, Crime, and Health. The total number of articles in the final dataset after duplicate removal was 18,876 articles which were all cleaned.

V.I Summary of the Dataset

• Total articles: 20,514
• Cleaned dataset: 18,876
• Average length: 350–700 words
• Duplicate reduction: 8%

V.II Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	82.4%	81.7%	80.9%	81.2%
Logistic Regression	85.3%	84.9%	84.3%	84.6%
SVM	89.1%	89.0%	88.4%	88.7%
Random Forest	86.7%	86.5%	86.2%	86.3%
LSTM	91.5%	91.2%	90.7%	90.9%
BiLSTM	93.4%	93.1%	92.6%	92.8%
BERT	96.2%	96.1%	95.8%	95.9%

Table 1 Model Performance

Key Findings:
• SVM was the most powerful traditional model.
• LSTM and BiLSTM used sequence modeling which in turn gave a boost to the performance.
• BERT provided the highest accuracy thus confirming that contextual embeddings are indispensable.
• Preprocessing has increased the final accuracy by more than 14%.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

V.III Confusion Matrix Insights

• Politics and Business were intertwined because of the economic-political issues that became major topics.
• The least incorrect category predicted was Sports, which was basically a consequence of using particular vocabulary of the domain.
• Crime and Health got the moderate overlap that was caused by the similar reporting patterns (cases, incidents).

V.IV Impact of Preprocessing

• Preprocessing has raised the maximum accuracy of the best model from 82% to 96.2%.
• The removal of duplicates and the normalization of text have greatly increased the trustworthiness of the classifier.

VI. DISCUSSION

This research has validated that the combination of automated scraping and sophisticated natural language processing (NLP) produces a drastic increase in the efficiency and scalability of news categorization systems. The article scraper was capable of wonderfully extracting thousands of articles, and then the transformer-based models allowed almost human-level accuracy in classification.

The advantages of the pipeline that was integrated are:

• Always-on and unlimited data gathering
• Lowered human effort
• Possibility of instant classification
• News domains that are heterogeneous supported

Despite the fact that there might be issues such as website variability, legal impediments, and computational cost, they are still doable with modern automation techniques

VII. APPLICATIONS

The application of this system can be seen in the following fields:

• News aggregation platforms
• Automated journalism
• Media monitoring tools
• Stock market & trend analysis
• Fake news detection
• Social media intelligence
• Academic linguistic research

VIII. CHALLENGES

The main challenges comprise of:

• HTML structure differences
• Legal issues related to web scraping
• Systems detection for humans only
• Transformers requiring substantial resources
• Unclean data from scraping
• Changes in website layout made often

From these, it is clear that adaptive scraping and model deployment with high efficiency are important.

IX. FUTURE WORK

The system can be made more effective by:

• Implementing a classification feature for multiple languages

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

- Joining together the functionalities of summarization and keyword extraction
- Employing AI for detecting DOM structure
- Scaling via cloud to achieve quicker scraping
- Collaborating with GPT-like generative models
- Real-time news streams inclusion

X. CONCLUSION

The comprehensive automation process for news scraping and classification system prepared the main contribution of this research. The pipeline worked in the background to collect more than 20,000 articles while also providing excellent classification accuracy, with BERT reaching 96.2%. The presented results guarantee that the combination of automation and cutting-edge NLP can considerably enlarge news analysis on a large scale. The developed system is a powerful base for the upcoming intelligent media processing applications that need to be real-time.

REFERENCES

[1]. Maududie, A., Retnani, W. E. Y., & Rohim, M. A. (2018, November). An approach of web scraping on news website based on regular expression. In 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT) (pp. 203-207). IEEE.
[2]. Sundaramoorthy, K., Durga, R., & Nagadarshini, S. (2017, April). Newsone—an aggregation system for news using web scraping method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140). IEEE.
[3]. Sundaramoorthy, K., Durga, R., & Nagadarshini, S. (2017, April). Newsone—an aggregation system for news using web scraping method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140). IEEE.
[4]. Bhujbal, M., Bibawanekar, M. B., & Deshmukh, P. (2023). News aggregation using web scraping news portals. International Journal of Advanced Research in Science, Communication and Technology, 3(2), 275-284.
[5]. Bhujbal, M., Bibawanekar, M. B., & Deshmukh, P. (2023). News aggregation using web scraping news portals. International Journal of Advanced Research in Science, Communication and Technology, 3(2), 275-284.
[6]. Thota, P., & Ramez, E. (2021, June). Web scraping of covid-19 news stories to create datasets for sentiment and emotion analysis. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (pp. 306-314)