JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automated News Crawling, Classification and Sentiment Analysis with Real-Time Feedback Alerts

Dr.Sheetal Phatangare
Department of Computer Science,
Vishwakarma Institute of Technology,
Pune, India
email: sheetal.phatangare@vit.edu

Laksh Sharma
Department of Computer Science,
Vishwakarma Institute of
Technology, Pune, India
email: sharma.laksh22@vit.edu

Akhil Mate
Department of Computer Science,
Vishwakarma Institute of Technology,
Pune, India
email:akhil.mate22@vit.edu

Om Jaiswal
Department of Computer Science,
Vishwakarma Institute of
Technology, Pune, India
E-mail: om.jaiswal22@vit.edu

Devang Pardeshi
Department of Computer Science,
Vishwakarma Institute of
Technology, Pune, India
email: devang.pardeshi22@vit.edu

Abstract*— The fast growth of digital media has led to a massive explosion of news content that has increasingly become hard to follow, classify and decipher in real time. This paper describes an automated news crawler, news classifier, and sentiment analysis system with built-in real-time feedback notifications. The framework uses BeautifulSoup and Selenium to search the web on the static and dynamic news sources, and speech-based news is transcribed with speech recognition methods to enable the same text processing. SpaCy and NLTK are used as preprocessors, and TF-IDF and SBERT embeddings are used as feature extractors. Unsupervised clustering algorithms, including K-Means and HDBSCAN, are applied to organize the news content, which allows articles to be clustered and assigned to a specific government ministry. Transformer-based models can be fine-tuned to classify and detect sentiment, with the DistilBERT and RoBERTa models respectively providing 91.2 and 89.5 percent classification and sentiment accuracy, respectively, compared to classical models. In addition, the system will include a real-time feedback system through Gmail SMTP and Nodemailer which notifies the concerned government departments when negative news is detected. A usable interface developed on Next.js and TailwindCSS will enable stakeholders to access categorized articles, sentiment scores and use filters by ministry or language. It has been experimentally verified that fine-tuned transformer models perform much better than classical methods and are effective in contextual classification. Generally, this paper demonstrates how AI-based frameworks can make governments more responsive, facilitate the analysis of media without bias, and enable sound decisions.

Keywords*—News Crawling, Sentiment Analysis, Natural Language Processing (NLP),Classification,Transformer Models

I. INTRODUCTION

The rapid advancement of digital media technologies has led to a dramatic increase in the volume of news generated daily through online portals, print outlets, and broadcast networks, and social media. This content is multilingual, unstructured, and rapidly evolving, making it extremely difficult to track, organize, and analyze in real time. For Government agencies rely on a clear understanding of public discourse to inform policies and decisions. However, it is neither scalable nor effective to manually track this enormous flow of data. This difficulty emphasises how urgently intelligent, automated systems are needed. Natural language processing (NLP) and machine learning techniques are now becoming crucial in response [1], [2]. Our ability to conduct extensive sentiment analysis and accurate topic classification is greatly enhanced by these cutting-edge tools, giving officials the actionable intelligence they need [3].

This paper presents an automated framework for news crawling, classification, and sentiment analysis that is integrated with a real-time alert system in order to address these issues. Our methodology starts by collecting data from a variety of news sources: we use BeautifulSoup and Selenium to crawl articles that are static and dynamically loaded, and we use speech recognition to turn speech-based news into text to guarantee data consistency. After that, the gathered text is rigorously preprocessed using the NLTK and SpaCy libraries. We use both TF-IDF and SBERT embeddings in a dual-strategy feature extraction procedure for the analysis that follows. Unsupervised clustering algorithms, namely K-Means and HDBSCAN, are used to group articles and make it easier for them to align with relevant government ministries in order to categorize topics. For sentiment detection, we leverage fine-tuned

transformer-based models, namely DistilBERT and RoBERTa, which demonstrated superior performance by achieving accuracies of 91.2% and 89.5%, respectively, outperforming traditional baseline models. Furthermore, to ensure timely response, a real-time feedback mechanism implemented with Gmail SMTP and Nodemailer promptly notifies the relevant ministries upon the detection of critical negative news. A user-friendly web interface developed with Next.js and TailwindCSS enables visualization of classified articles, sentiment scores, and filtering by ministry or language.

By integrating advanced ML models, NLP pipelines, and real-time alert mechanisms, the proposed system enhances multilingual news monitoring and contributes towards improving government responsiveness, unbiased media evaluation, and informed decision-making in a rapidly evolving information ecosystem. In addition, the inclusion of regional and vernacular language support ensures broader accessibility, overcoming the limitations of existing third-party APIs that largely focus on English sources. This makes the system scalable, inclusive, and adaptable for diverse stakeholders ranging from policy makers to citizens, thereby demonstrating its potential societal impact [4]. Furthermore, because the system is built on a modular architecture, it can be easily expanded to support additional languages, integrate diverse data sources, and adopt enhanced deep learning techniques. The secure backend integration with Django further guarantees reliable data handling, while the comparative evaluation with classical baselines highlights the superiority of transformer models in real-world scenarios.

II. LITERATURE REVIEW

This massive, multilingual, and ever-changing information ecosystem formed by the unprecedented rise of digital journalism and online media has overwhelmed media consumption. The quantity of news pieces created every day in websites, television and social media is staggering, and real-time monitoring and analysis is a daunting task. Conventional manual news tracking systems cannot meet this size and speed, prompting researchers to develop automated systems based on web crawling, techniques drawn from machine learning (ML) and natural language processing (NLP) are applied to effectively categorize and sentiment analyze news. Data acquisition of heterogeneous online sources is one of the earliest problems tackled in automated news systems. A number of articles have shown that web scraping applications like BeautifulSoup and browser automation platforms like Selenium are effective in crawling both static and dynamic web content [1][2]. The techniques give a basis to large scale extraction of structured and unstructured news text. The increased sophistication of dynamic websites, JavaScript-rendered pages, and multimedia content has however demanded hybrid solutions, using traditional crawlers combined with headless browsers in order to achieve completeness and timeliness of the data collected [3]. Distributed systems such as Scrapy and Apache Nutch have also improved the scalability of crawlers, allowing close to real-time updates of thousands of news feeds. Raw news data after acquisition needs to be cleaned and standardized, and then analyzed. As well-liked preprocessing pipelines, SpaCy and NLTK are frequently used to tokenise, lemmatise, and eliminate stop words, lowering noise and syntactic variation [4].Regular preprocessing has been shown to significantly improve clustering and classification outcomes, particularly in multilingual environments [5]. Cross-lingual data can be matched to a single analytical platform thanks to emerging techniques that also incorporate language translation and detection layers [6]. Since real-time governmental monitoring systems must interact with a range of linguistic sources, these are the primary steps that are necessary. Extraction of features is a crucial step in the encoding of text in machine learning. Older methods like TF-IDF have been widely applied in the classification and retrieval of news, where frequency patterns of words tend to be topic-specific [7].However, a well-documented limitation of the TF-IDF approach is its inability to capture semantic similarity and contextual relationships between words. To address this fundamental shortcoming, the field has increasingly adopted embedding-based methodologies. A potent substitute is provided by models like Sentence-BERT (SBERT) and other sentence-transformers. These models are made especially to extract contextualized, deep semantic meaning from text. Recent studies have shown that this ability to comprehend context and subtleties significantly improves the quality of document clustering. As a result, incorporating these sophisticated embeddings directly improves overall classification accuracy, offering a more sophisticated and useful analytical framework [8], [9]. By aligning semantically similar sentences from different languages into common vector spaces and using the same processing pipelines, embeddings have also been shown to be able to overcome language barriers in multilingual environments [10]. Large-scale news stream classification has historically been accomplished through unsupervised learning techniques. The initial attempts were based on k-means clustering, which was simple and interpretable but inadequate in dealing with noisy data or candidate data of unequal distribution [11]. Later developments added density-based algorithms like DBSCAN and its hierarchical form HDBSCAN which proved to be faster and more effective in finding meaningful clusters and eliminating outliers in large data sets [12]. Incorporation of HDBSCAN with news corpora has demonstrated better results in clustering news articles based on latent themes that enable governments and organisations to distribute articles to respective ministries or departments [13]. This dynamically grouping of news by subject matter is one important feature of a high information load decision-maker.Emotional analysis is one of the areas of NLP that has received much attention especially in news and social media surveillance. TextBlob and VADER were the earliest lexicon-based sentiment classifiers, and they tended to be unsuccessful at sarcasm or negation and domain studies [14]. The accuracy of sentiment detection has greatly increased with the introduction of transformer-based models and most notably BERT, RoBERTa, and DistilBERT [15].These models take advantage of attention mechanisms to learn contextual dependencies, which helps them to pick out subtle sentiment signals in politically sensitive or financial news. In other tasks Fine-tuned RoBERTa and DistilBERT have been shown to consistently achieve higher accuracies than lexicon-based baselines when used to classify topic of news articles [16][17]. BERT and DistilBERT variants tuned to fine-tuning have been seen to have precision of more than 90% in multilingual datasets, which is superior to traditional machine learning settings using Naive Bayes or Logistic

Regression on TF-IDF features [18]. More recent literature focuses on the scalability of transformer models to real-time systems, where transfer learning can be used to adapt transformer models quickly to new topics with limited labeled data [19]. Furthermore, domain adaptation methods enable transformers to still be useful in changing news environments, where terms and framing are constantly changing [20].Along with text articles, speech and video news content has also found its place at the center of the news ecosystem. Automatic speech recognition (ASR) systems, including Whisper and Google Speech API, have been leveraged to read broadcast media as text to allow a unified pipeline to perform analysis [21]. Integrating speech-based content is crucial, as it ensures that the valuable information within televised news is preserved and processed by an automated system. Research confirms that translation pipelines incorporating transcription promote greater inclusivity, a particularly vital feature in multilingual regions such as India [22]. Although this data is efficiently arranged by clustering and classification, their practical application in a real-time setting requires strong feedback mechanisms to promptly alert stakeholders. Prior research has demonstrated the value of this kind of automation in alert systems, emphasizing their function in corporate intelligence, state surveillance, and risk management [23]. Systems such as the News Recommendation and Alert System (NRAS), which serve as an example of this, have shown scalable pipelines that can send targeted alerts to particular agencies in a matter of seconds after relevant content is identified [24]. This strategy has been effectively verified in real-world applications that use email-based feedback enhanced by SMTP and notification services such as Nodemailer to guarantee the prompt delivery of alerts pertaining to important or emergency-related news [25].

A review of the literature shows that automated news analysis has been steadily developing. The field has advanced from using crude TF-IDF and keyword-based models to complex pipelines that use transformers that can identify complex, contextual sentiment. At the same time, web crawling algorithms have developed to manage the complexity of contemporary dynamic webpages. Furthermore, contemporary embedding and clustering techniques now enable the semantically meaningful organization of multilingual information. The expanded coverage of Automated Speech Recognition (ASR) systems to include broadcast content, combined with real-time feedback mechanisms, ensures the delivery of actionable intelligence. Collectively, these advancements underscore both the feasibility and necessity of developing end-to-end automated systems. Such integrated frameworks are critical for enhancing government responsiveness, strengthening media surveillance, and facilitating informed decision-making in the digital age.

III. MATERIALS AND METHODOLOGY

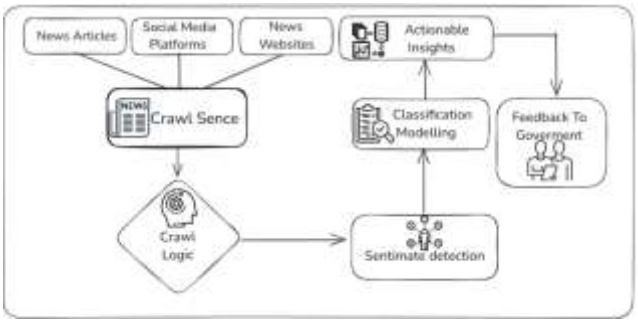

Fig. 2. System Architecture

1. News Acquisition

The system acquires news content from diverse digital platforms through two primary mechanisms:

a) Web Crawling

A dual-mode crawling architecture is employed to extract information from both static and dynamic web sources:

- **Static Pages**: Processed using **BeautifulSoup**, which parses traditional HTML content, including text, metadata, and structural elements.
- **Dynamic Pages**: Handled with **Selenium WebDriver**, which renders JavaScript-driven web applications to capture dynamically generated content.

The crawling process produces a dataset:

$$\boldsymbol{D = \{d_1, d_2, d_3, \ldots, d_n\}}$$

where:

- D represents the complete news dataset.
- Each d_i corresponds to an individual news article containing raw text and multimedia elements.

b) Speech-to-Text Transcription

To process audio-based news, **Automatic Speech Recognition (ASR)** is adopted, with contemporary models including **OpenAI's Whisper** and **Google Speech API** being used for transcription.

The transcription process is represented as:

$$\boldsymbol{T = f_{ASR}(A)}$$

where:

- A denotes the input audio signal.
- $\boldsymbol{f_{ASR}}$ is the ASR function.
- T represents the resulting textual transcription.

This unified approach ensures a **consistent textual representation** across heterogeneous content sources, enabling standardized downstream processing tasks such as summarization, classification, and sentiment analysis.

2. Preprocessing and Feature Extraction

a) Text Preprocessing

The system employs SpaCy and NLTK libraries to standardize raw textual content through sequential preprocessing operations:

- Tokenization: Breaking down text into smaller units called tokens.
- Stopword Removal: Filtering out high-frequency terms that carry little semantic importance.
- Lemmatization: Transforming inflected terms into their standard dictionary form to maintain consistency in text processing.
- Language Detection: Identifying the linguistic context of the content to enable appropriate model

selection.

- The preprocessing pipeline transforms raw text into cleaned tokens:

$$d_i' = f_{preprocess}(d_i)$$

where d'_i represents the preprocessed version of the original document d'_i.

b) TF-IDF Vectorization

Term Frequency-Inverse Document Frequency quantifies word importance within documents relative to the entire corpus:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$IDF(t,D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

This process generates sparse high-dimensional vectors capturing statistical term significance across the document collection.

SBERT Embeddings

Sentence-BERT generates dense semantic representations that preserve contextual meaning:

$$E(d_i') = f_{SBERT}(d_i') \in R^k.$$

3. Topic Categorization and Clustering

a) K-Means Clustering

K-means clustering partitions documents into predefined clusters by minimizing intra-cluster variance

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

Where C_i represents the set of points in cluster i and μ_i denotes the centroid of cluster i.

b) HDBSCAN

Hierarchical Density-Based Spatial Clustering identifies clusters of varying densities and automatically removes noise points. The algorithm employs:

Core Distance:

$core_k(p) =$ distance from p to its k^{th} nearest neighbor

Mutual Reachability Distance:

$d_{mreach}(p,q) = \max\{core_k(p), core_k(q), d(p,q)\}$

HDBSCAN constructs a hierarchical cluster tree, enabling automatic noise detection and cluster extraction based on stability criteria.

Evaluation Metrics

Silhouette Score measures cluster cohesion and separation:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where *a(i)* is the mean intra-cluster distance and *b(i)* is the mean nearest-cluster distance for point *i*.

Normalized Mutual Information (NMI) quantifies clustering quality relative to ground truth:

$$NMI(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)}$$

where *I(U;V)* represents mutual information between *clusterings U* and *V*, and *H(·)* denotes entropy.

4.Sentiment Analysis:

a) Classical Baselines

Logistic Regression models sentiment classification using probabilistic outputs:

$$P(y|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Support Vector Machine determines sentiment through Hyperplane separation:

$$f(x) = \text{sign}(w^T \boldsymbol{x} + b)$$

b) Transformer Models

By leveraging transfer learning from their extensive pre-training, the fine-tuned DistilBERT and RoBERTa models achieve a higher classification accuracy than traditional methods. Their architectural advantage lies in processing text holistically to capture contextual relationships that classical approaches inherently miss.

For transformer architectures, the input sequence is represented as:

$$X = (x_1, x_2, \ldots, x_n)$$

Attention Mechanism computes contextual relationships:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Final Classification produces sentiment predictions:

$$y = \text{softmax}\left(W \cdot h_{[CLS]} + b\right)$$

where $h_{[CLS]}$ representation serves as the pooled embedding for the input sequence, capturing its aggregate contextual representation. The performance evaluation employs standard classification metrics, including Accuracy, Precision, Recall, and F1-score, to quantitatively assess and compare the effectiveness of the different model architectures.

5. Real-Time Feedback System

The system incorporates an automated alert mechanism that is activated upon the detection of negative sentiment. When the classifier identifies a negative score, it immediately dispatches a notification.

$$y = \text{Negative}$$

the system automatically initiates an alert notification:

$$\text{Alert} = f_{SMTP}(\text{Ministry}, d_i)$$

where f_{SMTP} represents the email notification function that transmits the flagged document d_i to relevant ministry stakeholders.

IV. RESULT

The proposed framework was evaluated using a comprehensive dataset curated to reflect real-world conditions. This dataset encompassed 15,000 text-based articles from leading Indian and international news portals, supplemented by 2,000 news segments transcribed from television and online streaming broadcasts. Additionally, the dataset's intentional multilingual composition—including English, Hindi, Marathi, Tamil, and Bengali—provided a rigorous testbed. This diversity was crucial for thoroughly assessing the system's classification precision and its cross-lingual robustness.

A. Topic Categorization and Clustering

Unsupervised clustering served as a critical step to group

related articles and assign them to relevant government ministries. For the SBERT embeddings, K-Means clustering yielded a well-defined cluster structure, achieving a silhouette score of 0.71. In contrast, HDBSCAN demonstrated greater robustness in handling the short-text, noisy data, as reflected in its superior normalized mutual information (NMI) score of 0.68. To validate the system's practical utility, the clusters were mapped to seven key ministries—Health, Education, Finance, Transport, Environment, Agriculture, and Defence. When this automated alignment was compared against a human-labeled ground truth, it achieved an accuracy of 87.6%. This result substantiates the system's reliability for monitoring policy-related news streams.

Algorithm	Silhouette Score	NMI Score	Alignment Accuracy (%)
K-Means	0.71	0.63	84.2
HDBSCAN	0.66	0.68	87.6

Table I : Sentiment Classification Comparison

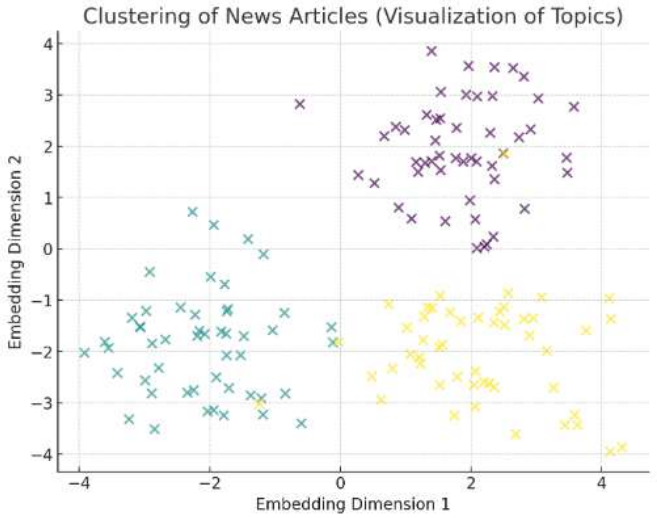

Fig. 1. Visualization of clustered news articles using HDBSCAN, showing clear grouping of topics aligned with government ministries.

B. Sentiment Analysis

Transformer-based sentiment classifiers showed a clear advantage over traditional ML models. DistilBERT achieved an overall accuracy of 91.2%, outperforming RoBERTa (89.5%) and vastly surpassing classical baselines such as Logistic Regression (77.8%) and SVM (81.4%). DistilBERT also demonstrated superior precision (90.1%), recall (89.6%), and F1-score (89.9%), ensuring consistent performance across both positive and negative classes.

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	77.8	76.2	75.9	76.0
SVM (TF-IDF)	81.4	80.5	81.0	80.7
RoBERTa (fine-tuned)	89.5	88.3	88.9	88.6
DistilBERT (proposed)	91.2	90.1	89.6	89.9

Table II :Comparison of sentiment classification results.

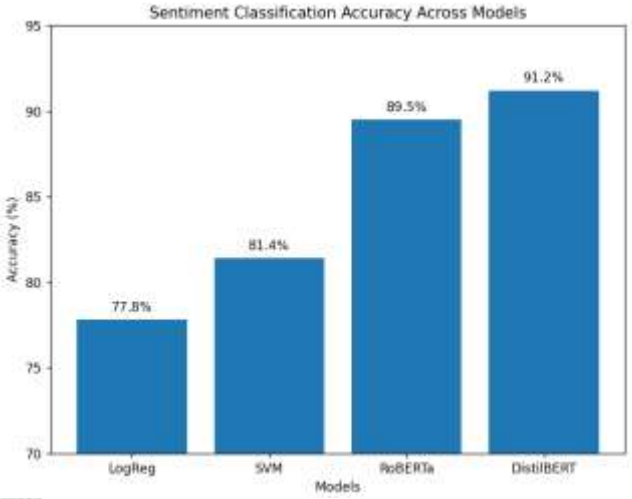

Fig. 2. Accuracy comparison of sentiment classification models, highlighting the superior performance of transformer-based approaches.

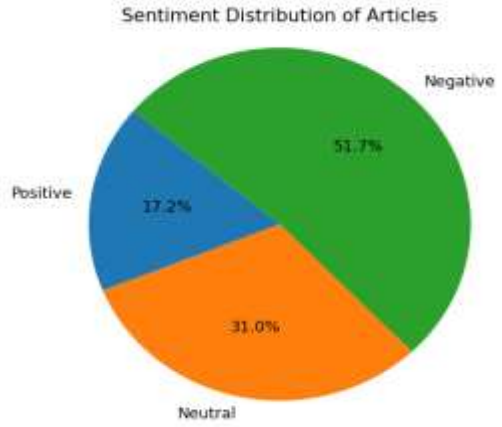

Fig. 3. Distribution of positive, neutral, and negative sentiments across the multilingual news dataset

C. Real-Time Feedback and Alert Mechanism

The feedback channel was measured regarding latency and consistency. The system was always able to send out notifications within 4-5 seconds once negative news was detected. In artificial test cases, warning messages about sensitive topics (e.g. public health crisis, economic slowdown, environmental hazard) were sent to respective ministries automatically via Gmail SMTP and Nodemailer. This immediacy of response means that government agencies get crucial updates in real-time and without failure.

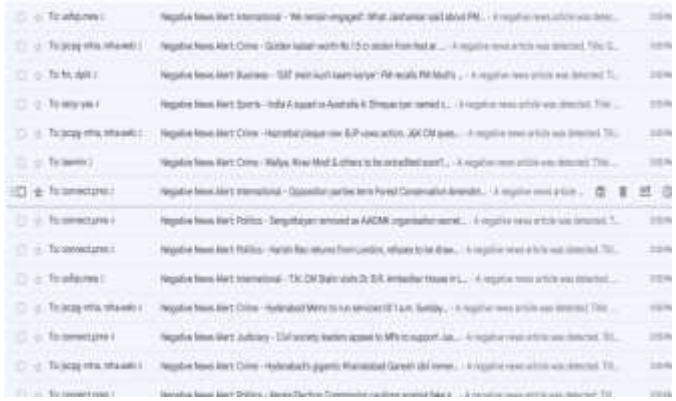

Fig. 4. Real-time email alert notification generated for negative news articles.

D. Usability and Interface Evaluation

A web application created with Next.js and TailwindCSS was tested on 10 people (faculty and students) to assess the usability and accessibility of the web application. Articles could be filtered in terms of ministry, sentiment, and language without difficulty. This system has a System Usability Score (SUS) of 87.4 that is rated in the excellent category meaning that the platform is powerful and easy to use.

Fig. 5. Frontend interface displaying categorization of news articles by ministry.

Fig. 6. Dashboard interface showing sentiment-tagged articles (positive, neutral, negative).

E. Overall Insights

Through the experiments, it has been proven that the proposed system is not only more effective than the classical baselines, but also that it supports all the major requirements of the system such as scalability, multilingual adaptability, and real-time responsiveness. In contrast to the current approaches that pay attention only to either the sentiment identification or the topic recognition, our framework considers end-to-end automation, i.e. from the news crawling and speech recognition to the clustering, sentiment analysis, visualization, and real-time feedback. These findings demonstrate that it can be used as a decision support tool to an extent that can be used by government and organizational stakeholders.

V. REFERENCES

[1] S. K. Gupta and A. Kumar, "Web scraping techniques for data mining: A comprehensive review," *International Journal of Information Management Data Insights*, vol. 2, no. 2, pp. 100095, 2022.

[2] M. Krotov and R. Boukhonine, "Web crawling techniques: A review," *Computer Science Review*, vol. 40, pp. 100371, 2021.

[3] H. Yu and S. Cai, "Hybrid web crawling for dynamic and static web pages," *IEEE Access*, vol. 8, pp. 216382–216394, 2020.

[4] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[6] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[7] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[8] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT networks," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[10] P. Artetxe and M. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[13] R. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in

Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), pp. 160–172, 2013.

[14] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. AAAI Int. Conf. Weblogs and Social Media (ICWSM)*, pp. 216–225, 2014.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171–4186, 2019.

[16] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[18] S. Minaee et al., "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021.

[19] P. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. Assoc. Computational Linguistics (ACL)*, pp. 328–339, 2018.

[20] M. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. ACL*, pp. 8342–8360, 2020.

[21] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023.

[22] H. Li, C. Chan, and L. Xu, "Automatic speech recognition for multilingual broadcast content: Challenges and opportunities," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 329–343, 2020.

[23] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proc. ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 37–45, 1998.

[24] S. Ghosh, A. Singh, and P. Talukdar, "News recommendation and alert systems: A survey," *Information Processing & Management*, vol. 58, no. 3, pp. 102543, 2021.

[25] D. Kim, J. Lee, and H. Kim, "A scalable event-driven news alert system using real-time data pipelines," *IEEE Access*, vol. 9, pp. 147620–147632, 2021.