Preprints.org

MIND-SBERT: A Text Analytics Platform for Transparent Insight Generation and Evidence Synthesis from News Data

[Ahmad Farooq](#) [*]

Article

MIND-SBERT: A Text Analytics Platform for Transparent Insight Generation and Evidence Synthesis from News Data

Ahmad Farooq

University of Arkansas at Little Rock; afarooq@ualr.edu

Abstract

This paper presents MIND-SBERT, an innovative text analytics platform that transforms large-scale unstructured news data into actionable insights through an integrated analytical pipeline. The platform leverages the Microsoft News Dataset (MIND) and advanced natural language processing techniques to create a comprehensive analytics solution focused on transparency and trust. By combining Sentence-BERT (SBERT) for semantic analysis, GPT-4o for content synthesis, and a multi-dimensional analytical evaluation framework, MIND-SBERT provides a powerful tool for data-driven decision support. The platform incorporates detailed analytical performance metrics including quantitative measures (ROUGE, BLEU, METEOR) and qualitative dimensions (conciseness, relevance, coherence, accuracy, fairness) for comprehensive output assessment. A critical innovation is the trust auditing mechanism that identifies and corrects potential inaccuracies in analytical outputs. The interactive Streamlit-based analytics dashboard enables users to explore data patterns, analyze multi-document insights with transparent evaluation metrics, and understand the analytical process through detailed explanations. This human-centered analytics approach provides users with deep visibility into the platform's analytical decision-making, fostering trust and enabling informed use of generated insights for evidence-based decision support.

Keywords: text analytics; trust in AI analytics; human-centered analytics; evidence synthesis; analytical decision support; natural language processing; MIND dataset

---

1. Introduction

The exponential growth of unstructured digital data in the 21st century has created critical analytical challenges for organizations seeking to extract actionable insights from vast information repositories. In the domain of news and current events, where data streams are continuously updated and expanded, the ability to perform rapid analytics on textual content is particularly crucial for informed decision-making. This paper introduces MIND-SBERT, a comprehensive text analytics platform that addresses these challenges by integrating advanced natural language processing (NLP) techniques within a transparent analytical pipeline, with a primary focus on enhancing trust and explainability in the analytics process.

The MIND-SBERT platform is built upon the Microsoft News Dataset (MIND), a large-scale English news dataset that provides a rich corpus for developing and evaluating text analytics methodologies [1]. This dataset offers diverse news content across multiple categories, making it an ideal foundation for demonstrating advanced analytics capabilities on unstructured text data. By combining this dataset with state-of-the-art analytical models and evaluation techniques, our platform offers a powerful solution for transparent data analysis and insight generation.

The field of text analytics has evolved significantly from simple keyword-based analysis to sophisticated semantic understanding and synthesis capabilities. Traditional analytical approaches often struggle with capturing contextual relationships and generating meaningful insights from

unstructured data. Modern text analytics, leveraging deep learning and semantic embeddings, enables more nuanced understanding and analysis of textual content [2]. This evolution is particularly valuable in news analytics, where understanding complex relationships between information pieces is crucial for comprehensive analysis.

Simultaneously, the need for trustworthy and explainable analytics has grown as organizations increasingly rely on AI-driven insights for critical decisions. The opacity of many analytical systems can lead to mistrust and misuse of generated insights. To address this challenge, MIND-SBERT incorporates multiple features designed to enhance analytical transparency and build user trust through comprehensive evaluation and explanation mechanisms.

The primary objectives of this research are:

1. To develop an efficient text analytics pipeline using semantic analysis techniques, specifically leveraging Sentence-BERT (SBERT) for understanding contextual relationships in unstructured data.
2. To implement an advanced content synthesis capability that generates coherent analytical summaries from multiple data sources, using state-of-the-art language models while providing clear explanations of the analytical process.
3. To establish a comprehensive analytical evaluation framework that assesses output quality through multiple quantitative and qualitative dimensions, supporting human judgment in analytical decision-making.
4. To implement trust auditing mechanisms that identify and correct potential inaccuracies in analytical outputs, thereby enhancing the reliability of the analytics platform.
5. To create an interactive analytics dashboard that enables users to explore data patterns, understand analytical processes, and make informed decisions based on transparent insights.

To achieve these objectives, we employ Sentence-BERT (SBERT) for semantic analysis [4], which has demonstrated exceptional performance in capturing semantic relationships in text data. For content synthesis, we utilize GPT-4o, a powerful language model capable of generating coherent analytical summaries. The evaluation framework incorporates established metrics including ROUGE [5], BLEU [6], and METEOR [7], alongside novel semantic similarity measures based on SBERT embeddings.

A key innovation of our platform is the integration of a human-centered analytics evaluation framework. This framework assesses analytical outputs based on nine qualitative dimensions: conciseness, relevance, coherence, accuracy, non-redundancy, readability, fairness, consistency, and resilience to input noise. By leveraging GPT-4o for generating these qualitative assessments, we provide nuanced, human-interpretable evaluation of analytical quality that complements quantitative metrics.

Furthermore, MIND-SBERT addresses critical challenges in trust and reliability through its analytical auditing capabilities. The platform implements mechanisms to detect and correct potential inaccuracies in generated insights, commonly referred to as "hallucinations" in generative AI systems [8]. This feature enhances the trustworthiness of the analytics platform for mission-critical applications.

The integration of these advanced analytics techniques, comprehensive evaluation metrics, and user-centered design positions MIND-SBERT as a powerful platform for navigating complex unstructured data landscapes. By enabling efficient analysis and synthesis of relevant information while providing transparent explanations and evaluations, our platform has the potential to significantly enhance data-driven decision-making across various domains, from business intelligence and market analysis to policy research and strategic planning.

The remainder of this paper is organized as follows: Section 2 describes the analytical framework and methods, including the data processing pipeline, semantic analysis techniques, content synthesis approach, evaluation metrics, and user interface design. Section 3 presents empirical results demonstrating the platform's analytical capabilities, evaluation performance, and trust mechanisms. Section 4 discusses the implications of our findings, the platform's strengths and limitations, and potential

areas for future enhancement. Finally, Section 5 concludes by summarizing the main contributions and the potential impact of MIND-SBERT on the field of transparent and trustworthy text analytics.

2. Materials and Methods

2.1. Analytical Framework and Pipeline Design

The MIND-SBERT platform implements a comprehensive analytical pipeline that transforms unstructured news data into actionable insights through four integrated stages:

1. **Data Integration Layer**: Processes raw news articles from the MIND dataset, extracting structured features (id, category, subcategory, title, abstract, URL, full text) and preparing them for analytical processing.
2. **Semantic Analysis Layer**: Employs SBERT to generate dense vector representations of textual content, enabling sophisticated pattern recognition and similarity analysis across the document corpus.
3. **Synthesis and Insight Generation Layer**: Utilizes GPT-4o to synthesize information from multiple sources, generating coherent analytical summaries that capture key insights and relationships.
4. **Evaluation and Trust Layer**: Implements multi-dimensional assessment of analytical outputs, including quantitative metrics, qualitative evaluation, and trust auditing to ensure reliability.

This pipeline design enables end-to-end analytics on unstructured text data while maintaining transparency at each processing stage. The modular architecture allows for independent optimization of each component while ensuring seamless integration for comprehensive analytical workflows.

2.2. Dataset and Data Preparation

The analytical platform utilizes the Microsoft News Dataset (MIND), specifically the training subset containing diverse news articles across multiple categories [1]. The dataset is structured in TSV format with rich metadata, providing an ideal corpus for demonstrating text analytics capabilities on real-world unstructured data.

2.3. Semantic Analysis Component

The semantic analysis component employs the all-MiniLM-L6-v2 model from Sentence-BERT (SBERT) [4]. This model was selected for its ability to generate semantically meaningful embeddings that capture contextual relationships in text data. The semantic analysis process involves:

1. Encoding user queries into dense vector representations
2. Generating embeddings for all articles in the corpus
3. Computing cosine similarity metrics to identify relevant content
4. Ranking and selecting top N articles based on semantic relevance

This approach enables the platform to perform sophisticated pattern matching and relationship discovery, supporting advanced analytical queries beyond simple keyword matching.

2.4. Content Synthesis and Insight Generation

The synthesis component implements two analytical approaches:

2.4.1. Single-Document Analysis

For individual articles, the platform generates analytical summaries using GPT-4o, prompted to create concise, coherent insights that capture essential information and relationships within the document.

2.4.2. Multi-Document Synthesis

For comprehensive analysis across multiple sources, a hierarchical approach is implemented:

1. Cluster related articles using K-means on SBERT embeddings

2. Generate cluster-level analytical summaries
3. Synthesize final insights by integrating cluster-level analyses

An iterative refinement process enhances analytical quality based on evaluation feedback, ensuring high-quality insight generation.

2.5. Analytical Evaluation Metrics

MIND-SBERT incorporates a comprehensive evaluation framework to assess analytical output quality:

- **Quantitative Metrics**: ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) for n-gram overlap analysis [5], BLEU score for precision assessment [6], METEOR score for semantic alignment [7]
- **Semantic Similarity**: Cosine similarity between SBERT embeddings of source and synthesized content
- **Qualitative Analytics**: GPT-4o-based evaluation across nine dimensions: conciseness, relevance, coherence, accuracy, non-redundancy, readability, fairness, consistency, and resilience
- **Trust Auditing**: Mechanisms to identify and correct potential inaccuracies in analytical outputs

This multi-faceted evaluation approach provides comprehensive assessment of analytical quality, supporting informed decision-making.

2.6. Interactive Analytics Dashboard

The platform features an interactive dashboard built with Streamlit, designed to support analytical workflows:

1. Query interface for analytical exploration
2. Transparent display of semantic analysis results with explanations
3. Individual and multi-document analytical summaries with comprehensive metrics
4. Trust indicators and accuracy assessments
5. Export capabilities for further analysis

This user-centered design enables analysts to effectively explore data patterns and derive insights while maintaining full visibility into the analytical process.

3. Results

The MIND-SBERT platform demonstrated robust analytical capabilities across multiple evaluation dimensions. Figure 1 illustrates the interactive analytics dashboard, showcasing the query interface and analytical results presentation.

Figure 1. Analytics dashboard interface showing query input and processing pipeline (left). Analytical results display with semantic relevance scores and transparency features (right).

3.1. Semantic Analysis Performance

For the analytical query `usa china relations`, the platform successfully identified and ranked relevant articles through semantic analysis. For example, the article "Is It Time For America To Begin

Decoupling From Communist China?" was retrieved with a semantic similarity score of 0.6270. The platform provides transparent explanations for analytical decisions:

```
This article was selected with a relevance score of 0.6270 because:
•   It belongs to the 'news' category, relevant for geopolitical analysis.
•   High semantic similarity indicates strong contextual alignment with the query.
•   The SBERT model identified significant thematic matches in the content.
```

These explanations enhance analytical transparency, enabling users to understand and validate the platform's analytical reasoning.

3.2. Analytical Output Quality

3.2.1. Single-Document Analytics

The platform generated high-quality analytical summaries for individual articles. Table 1 presents evaluation metrics for a representative analytical output. While traditional metrics show modest scores, the high semantic similarity (0.7568) indicates strong preservation of analytical meaning—a critical factor for abstractive analytics that these metrics may not fully capture [11].

Table 1. Scores (left) for the revised summary (right) of the retrieved article "Is It Time For America To Begin Decoupling From Communist China?"

Metric Score	Score
ROUGE-1	0.0768
ROUGE-2	0.0229
ROUGE-L	0.0470
BLEU	0.0000
METEOR	0.0221
Semantic Similarity	0.7568

Revised Summary with Hallucinations Removed or Corrected:

The article critiques the U.S. policy of supporting China's rise, arguing it has led to negative consequences for American values and interests. The author contends that rather than liberalizing China, this policy has caused a decline in American adherence to free speech and democratic principles, exemplified by the self-censorship of various businesses, such as the NBA, automobile, fashion, and lodging industries, to appease China. The article references Robert Spalding's book, which details how American organizations are manipulated by Chinese influence. Moreover, it highlights recent U.S. government actions against China, such as legal measures and policy proposals, to illustrate a shift towards a more confrontational stance.

Explanation of the Changes Made in the Revised Summary:

- Added mentions of other industries (automobiles, fashion, lodging) that have also practiced self-censorship for China to give a broader context.
- Clarified that Robert Spalding's book is one point among several in the article, which addresses various facets of Chinese influence.
- Included references to recent U.S. governmental actions and legislative considerations to fully reflect the actions discussed in the original article.

3.2.2. Multi-Document Analytics

Multi-document synthesis demonstrated the platform's capability to integrate insights across sources. The analytical output for "usa china relations" achieved a semantic similarity of 0.5777, indicating effective cross-document synthesis while maintaining thematic coherence.

Table 2. Scores (left) for the generated multi-document summary (right) of all retrieved articles.

Metric Score	Score
ROUGE-1	0.1027
ROUGE-2	0.0371
ROUGE-L	0.0515
BLEU	0.0000
METEOR	0.0328
Semantic Similarity	0.5777

Revised Summary:

The text highlights the increasing tensions between the USA and China, with significant attention given to critiques from American officials such as Secretary of State Mike Pompeo and Vice President Mike Pence against the Chinese Communist Party (CCP). These critiques involve accusations of opposing U.S. interests and human rights abuses, particularly in regions such as Xinjiang and Hong Kong. U.S. lawmakers are promoting legislation to sanction Hong Kong officials involved in human rights violations, reflecting a significant interest in Washington to challenge Beijing's policies on multiple fronts, including trade and human rights. Meanwhile, geopolitical tensions are exacerbated by the contentious issue of Taiwan, with Beijing asserting territorial claims over the island and opposing U.S. involvement that supports Taiwan's autonomy. From a commercial perspective, Robert Spalding's book discusses China's influence on American organizations, further emphasizing the adversarial relationship. Trade disputes remain a critical area of conflict, with ongoing tensions risking further escalation. Jack Ma, co-founder of Alibaba, warns that U.S.-China relations could face prolonged turbulence if trade issues are not resolved pragmatically. Despite these macroeconomic tensions, Alibaba's performance remains robust, driven by strong consumer spending in China. Overall, the text portrays a deteriorating relationship between the USA and China, manifested in economic, diplomatic, and political dimensions.

Explanation of the Changes Made in the Revised Summary:

- *Removed the Reference to Alibaba as a Direct Economic Threat:* Cleared up any inaccuracies by solely referring to Jack Ma's warning about potential prolonged turbulence without implying direct U.S. economic threats.
- *Adjusted the Bipartisan Resolve Statement:* Modified the language to avoid overgeneralizing the bipartisan resolve, focusing instead on specific actions taken by parts of the U.S. government.
- *Clarified Trade Disputes and Tariffs:* Removed the vague mention of "new U.S. tariffs" to make it clear that the focus is on ongoing trade tensions and potential risks of escalation.

3.3. Human-Centered Analytical Evaluation

The platform's qualitative evaluation framework provided nuanced assessment of analytical outputs. For the multi-document summary on USA-China relations, the evaluation yielded the following assessment, highlighting its strengths in relevance, accuracy, and fairness, while also identifying areas for improvement.

Conciseness (*effectively conveys important information while being brief*) 5/10

The summary is fairly detailed, including various aspects and quotes from different parts of the original text. While it captures key points, it could be more concise by focusing on the most significant aspects and reducing secondary details.

Relevance (*information is relevant to the main topic*) 9/10

The summary effectively captures the relevant information that revolves around U.S.-China tensions, economic conflicts, human rights issues, and geopolitical concerns. However, it could omit or condense some less central points to enhance focus.

Coherence (*clear structure and flow of ideas*):   7/10

The flow of ideas is mostly logical and structured, moving from critiques by U.S. officials to specific issues like Hong Kong, trade wars, and Taiwan. However, the narrative could be more streamlined to improve readability and flow.

Accuracy (*factually correct, no false or misleading information*)   8/10

Based on the original text, the summary appears factually accurate, with no apparent false or misleading information. It appropriately represents the points made in the original text.

Non-redundancy (*no repeated information*)   6/10

There are minimal issues with redundancy, but the summary could be trimmed further to avoid repetitive elements and ensure each point adds new information or perspective.

Readability (*easily understood by the target audience*)   7/10

The language is generally clear and comprehensible, but the density of the information could make it challenging for some readers. Simplifying some complex sentences and incorporating more transitions may aid readability.

Fairness (*objective and unbiased presentation*)   8/10

The summary presents an objective view, capturing multiple perspectives and avoiding biased language. However, a bit more balance in discussing both U.S. and Chinese positions could enhance impartiality.

Consistency (*consistent with the original in style and tone*)   7/10

The summary maintains a tone consistent with the original text but could slightly better match the nuanced style of the original by including specific authorial nuances and editorial notes.

Resilience to input noise (*accurate despite noisy input*)   6/10

The summary retains the core ideas despite the lengthy and detailed nature of the original text. Nevertheless, minor misinterpretations or missed nuances indicate room for improvement.

3.4. Trust Auditing Results

The trust auditing mechanism successfully identified and corrected three instances of potential inaccuracies in multi-document synthesis:

> 1. **Economic Entities Like Alibaba Citing Potential Threats to U.S. Economic Interests:**
>    ```
>    The original text mentions Jack Ma and the potential "turbulence" in
>    U.S.-China relations but does not specifically accuse Alibaba of posing a
>    threat to U.S. economic interests.
>    ```
> 2. **Bipartisan Resolve in U.S. Government to Confront Beijing:**
>    ```
>    The original text indicates a legislative interest and critiques from
>    officials like Pompeo, but it does not explicitly show an overarching
>    "bipartisan resolve" to confront Beijing on all described fronts (diplomacy,
>    trade, and human rights).
>    ```
> 3. **Statements about Trade Disputes and Potential for New Tariffs:**
>    ```
>    While the original text mentions ongoing trade disputes and a possibility of
>    further tariffs, the detail about "new U.S. tariffs" is vague and may imply
>    recent actions not directly supported by the text.
>    ```

The platform automatically generated corrected analytical outputs, demonstrating its commitment to accuracy and reliability in analytical processes.

3.5. Analytics Dashboard Functionality

The interactive dashboard successfully supported analytical workflows through:

- Intuitive query interface for analytical exploration
- Transparent presentation of semantic analysis results
- Comprehensive display of evaluation metrics with interpretive guidance
- Visual representation of qualitative assessment dimensions
- Clear communication of trust auditing results and corrections

These features collectively enable analysts to conduct thorough investigations while maintaining full visibility into the analytical process.

4. Discussion

The MIND-SBERT platform demonstrates significant advances in transparent text analytics, with implications for data-driven decision support across multiple domains.

4.1. Semantic Analysis for Advanced Analytics

The use of SBERT for semantic analysis enables sophisticated pattern recognition beyond traditional keyword-based approaches. The high semantic similarity scores between queries and retrieved content validate this approach for analytical applications. This capability is particularly valuable for discovering non-obvious relationships in large text corpora [10].

4.2. Synthesis Quality and Analytical Value

While traditional metrics (ROUGE, BLEU, METEOR) show modest scores, these measures were designed for extractive tasks and may not fully capture the value of abstractive analytics [9]. The high semantic similarity scores and positive qualitative evaluations suggest that the platform successfully preserves analytical meaning while generating coherent insights.

4.3. Comprehensive Evaluation as Analytics

The multi-dimensional evaluation framework itself represents an analytical contribution, providing decision-makers with nuanced understanding of output quality. This approach aligns with calls for more comprehensive evaluation in AI systems [12] and demonstrates how evaluation can be integrated as a core analytical capability.

4.4. Trust Through Analytical Transparency

The trust auditing mechanism addresses critical concerns about AI reliability in analytical applications [8]. By identifying and correcting potential inaccuracies, the platform demonstrates how transparency mechanisms can be embedded within analytical workflows to enhance trustworthiness.

4.5. Human-Centered Analytics Design

The interactive dashboard exemplifies human-centered analytics, providing analysts with tools to explore data while understanding the underlying analytical processes. This transparency aligns with principles of interpretable AI [13] and supports more effective human-AI collaboration in analytical tasks.

4.6. Limitations and Future Directions

While demonstrating strong capabilities, opportunities for enhancement include:

- Development of analytics-specific evaluation metrics that better capture insight quality
- Integration of domain-specific analytical models for specialized applications

- Enhanced visualization capabilities for complex analytical relationships
- Real-time analytics on streaming data sources
- Incorporation of user feedback for continuous analytical improvement

5. Conclusions

The MIND-SBERT platform represents a significant advancement in transparent text analytics, successfully integrating semantic analysis, content synthesis, and comprehensive evaluation within a unified analytical framework. Key contributions include:

1. Development of an end-to-end analytics pipeline that transforms unstructured text into actionable insights
2. Integration of semantic analysis capabilities enabling sophisticated pattern discovery
3. Implementation of multi-dimensional evaluation as a core analytical function
4. Trust auditing mechanisms that enhance reliability for critical applications
5. Human-centered design promoting transparency and informed decision-making

The platform demonstrates how advanced NLP techniques can be effectively integrated within analytical workflows while maintaining transparency and trust. As organizations increasingly rely on AI-driven analytics, systems like MIND-SBERT will play a crucial role in enabling reliable, explainable insights from complex unstructured data. Future work will focus on expanding analytical capabilities, improving evaluation metrics, and exploring domain-specific applications across business intelligence, policy analysis, and strategic planning contexts.

References

1. Wu, F.; Qiao, Y.; Chen, J.H.; Wu, C.; Qi, T.; Lian, J.; Liu, D.; Xie, X.; Gao, J.; Wu, W.; et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3597–3606.
2. Uyar, A.; Aliyu, F.M. Evaluating search features of Google Knowledge Graph and Bing Satori: entity types, list searches and query interfaces. *Online Information Review* **2015**, *39*, 197–213.
3. Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023* **2016**.
4. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* **2019**.
5. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*, 2004, pp. 74–81.
6. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
7. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
8. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* **2020**.
9. Liu, F.; Liu, Y. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, 2008, pp. 201–204.
10. Bai, X.; Wang, M.; Lee, I.; Yang, Z.; Kong, X.; Xia, F. Scientific paper recommendation: A survey. *IEEE Access* **2019**, *7*, 9324–9339.
11. Kryściński, W.; McCann, B.; Xiong, C.; Socher, R. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840* **2019**.

12. Celikyilmaz, A.; Clark, E.; Gao, J. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799* **2020**.
13. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.