Analysis of Civil Aviation Incident Reporting and Discovery Issues Based on Text Mining Techniques

Zhenzhong Zhang*
CAAC Academy, Civil Aviation Flight University of China, Guanghan, Sichuan, China
**Corresponding Author.*

Abstract: In the aviation industry, ensuring safe operations is a top priority, however, the traditional regulatory model often faces problems such as information asymmetry and unclear inspection focus, resulting in a low problem detection rate and inefficient supervision. With the rapid development of big data and artificial intelligence technology, the construction of a risk management early warning system through the construction of a smart supervision platform can analyze the problems of the supervised parties in a multi-dimensional way, analyze and assess the overall operational risks of airports and airlines, so as to realize accurate supervision. In this paper, we will discuss in detail how to use text technology to analyze the unsafe event reports and inspections to find problems, in order to help realize precise supervision.

Keyword: Aviation Safety; Text Mining; Natural Language Processing; Problem Categorization; Risk Warning

1. Introduction

In today's society, with the rapid development of science and technology and the full advent of the information age, all industries are experiencing unprecedented changes, and the field of aviation safety regulation is no exception [1]. Aviation safety, as the cornerstone of ensuring smooth air traffic, passenger life and property safety, is always at the forefront of industry management. However, traditional regulatory tools and methods are facing increasingly serious challenges: the explosive growth of information, the diversification of safety hazards and the complexity of risk identification have all put forward higher requirements for the efficiency and effectiveness of regulation. Against this background, how to apply advanced technologies and methods to enhance the intelligence level of supervision and realize accurate supervision has become a key issue to be solved in the field of aviation safety supervision [2].

The analysis of civil aviation incident reports and discovery problems based on text mining technology is a research topic that has arisen from the needs of this era. The purpose of this paper is to explore how to deeply analyze the massive unsafe event reports and complex inspection discoveries through text technology [3], especially Natural Language Processing (NLP) and text mining technology, in order to accurately capture the key information in the ocean of information, and to provide powerful support for aviation safety supervision [4]. The research in this paper is not only of great significance for improving the efficiency and quality of supervision, but also lays a solid foundation for realizing intelligent supervision and constructing a risk early warning system, which is a key step towards a preventive supervision model.

In the traditional mode of aviation supervision, problems such as information asymmetry and unclear inspection focus often lead to low supervisory effectiveness. The existence of these problems stems from the huge amount and complexity of regulatory information on the one hand, and the lack of effective technical means to efficiently process and deeply mine such information on the other. The rapid development of big data and artificial intelligence technology provides a new solution for this. Through the construction of a smart regulatory platform, it is possible to integrate data from multiple sources, utilize advanced analytical tools, conduct risk assessment and early warning of aviation operations from multiple dimensions, and realize accurate identification and timely intervention of problems in airports, airlines and other regulated parties.

The application of text technology, especially natural language processing and text mining technology, is the core driving force of this transformation [4]. Text data, as an important part of the information ocean, contains rich connotations and values, but in the regulatory field, how to extract key information from the huge amount of unsafe event reports and inspection findings has been a major challenge to enhance regulatory effectiveness. In this paper, we will discuss in depth how text technology can solve this challenge, including but not limited to automatic classification and clustering of events, rapid extraction of key factors, application of sentiment analysis, and how to realize in-depth profiling and trend prediction of problems through technical means such as intelligent comparison and knowledge graph.

Through the research in this paper, we expect to show the wide range of applications of text technology in the analysis of unsafe incident reports, such as how to automatically categorize the reports into different categories of safety issues by using event classification and clustering techniques to reveal the relationships and patterns among events; how to accurately identify the responsible person, equipment model, and other key information through key factor extraction techniques, such as Named Entity Recognition (NER) and key phrase extraction; and how to assess the social impact and public reaction through sentiment analysis to provide a more comprehensive perspective for regulatory decision-making. information; and how to assess the social impact and public reaction of an incident through sentiment analysis to provide a more comprehensive perspective for regulatory decision-making.

In addition, this paper will also analyze in depth the text analysis process of the inspection discovery issues, and introduce how to use text categorization, trend analysis, text summarization and information extraction, risk early warning and assessment, intelligent comparison mechanism, and knowledge graph construction, etc., to deeply mine the inspection reports, which not only improves the efficiency and precision of the analysis, but also provides a scientific basis for the development of more accurate and efficient regulatory strategies.

In summary, the research in this paper not only focuses on the innovative application of the technology level, but also focuses on how these technologies can serve the practical needs of aviation safety regulation, providing theoretical support and technical paths for realizing intelligent regulation [5]. Through the application of these cutting-edge technologies, the future aviation regulatory system will be able to operate more intelligently and efficiently, effectively respond to all kinds of safety challenges, ensure the stability and safety of the air transportation industry, and contribute to the safe and sustainable development of the global aviation industry.

2. Application of Text Technology in Precision Supervision

In today's era of information explosion, text data, as an important part of the data ocean, contains rich information value. Especially in the field of regulation, in the face of massive reports of unsafe events and complex inspection findings, how to efficiently and accurately extract key information from them has become a key challenge to enhance the efficiency and effectiveness of regulation. Text technology, especially Natural Language Processing (NLP) and text mining technology, provides a new perspective and tool for realizing accurate regulation through its powerful information extraction, understanding and analysis capabilities.

2.1 Textual Analysis of Insecurity Reports

As an important source for regulators to obtain information and assess risks, reports on unsafe incidents cover the causes and consequences of the incidents, the scope of impact and possible solutions. However, traditional manual reading and analysis methods can hardly cope with the processing needs of large-scale reports, which may easily lead to information omission or judgment bias. The application of text technology, on the other hand, can effectively solve this problem and realize the in-depth analysis and utilization of insecurity incident reports.

2.1.1 Event classification and clustering

Event categorization and clustering is the basic application of text technology in the analysis of unsafe event reports. With text categorization techniques, reports can be automatically assigned to predefined

categories, such as equipment failure, operational errors, environmental factors, etc. This process often relies on supervised learning methods such as Support Vector Machines (SVMs), logistic regression, etc., which train the model with labeled datasets, enabling it to accurately identify key themes in the reports [6]. In addition, unsupervised clustering techniques such as K-means, hierarchical clustering, etc., can further refine the grouping of events and reveal subtle differences and patterns between events. In particular, topic models such as LDA (Latent Dirichlet Allocation) automatically generate a series of topics and assign topic probabilities to each document through probabilistic statistical methods, thus revealing the deep structure and correlations behind events. The application of these techniques enables regulatory units to grasp the distribution of types of insecurity events on a macro level and to formulate regulatory policies and preventive measures in a targeted manner.

2.1.2 Key factor extraction

In incident reporting, accurate identification and extraction of key information is the basis for effective regulation. Named Entity Recognition (NER) technology can automatically identify entities in the text, such as names of people, organizations, equipment models, geographic locations, etc., which are crucial for tracking those responsible for the incident and analyzing the scope of impact of the incident. Meanwhile, key phrase extraction techniques, such as TF-IDF and TextRank, are able to extract the most representative and informative words and phrases from the text, helping regulators to quickly capture the core elements of an incident. Through the combined application of these technologies, regulators are able to quickly locate the source of the problem and provide a clear direction for subsequent investigation and treatment.

2.1.3 Sentiment analysis

Sentiment analysis techniques provide another dimension of insight for regulatory decision-making. It assesses the level of impact and public reaction to an unsafe event by analyzing the emotional coloring of the text, such as positive, negative, and neutral. In an unsafe incident report, a strong expression of negative sentiment may indicate the severity of the incident or a high level of public concern about it, which requires immediate response and prioritization by the regulatory unit. Sentiment analysis not only helps supervisors to quickly identify emergencies, but also provides clues for improving management systems and enhancing employees' safety awareness by analyzing implicit sentiments in reports, such as employees' dissatisfaction or concerns about the safety system.

2.2 Textual Analysis of Inspection Findings

In modern regulatory practice, inspection reports are not only an important basis for evaluating compliance and assessing risks, but also a key tool for promoting continuous improvement and ensuring safe operation. The application of text technology, especially Natural Language Processing (NLP) and information extraction technology, has shown great potential in improving the efficiency and precision of analyzing inspection findings. By exploring the rich information in inspection reports, regulators can realize more accurate and efficient regulatory strategies.

2.2.1 Problem classification and trend analysis

Traditional examination report analysis often relies on manual browsing and manual categorization, which is not only time-consuming but also error-prone. With the introduction of text categorization techniques, this process has been significantly optimized. Using advanced machine learning models, such as deep neural networks and random forests, problems in inspection reports, such as insufficient equipment maintenance, operating procedure violations, and substandard environmental conditions, can be automatically categorized [7]. Combined with time-series analysis, regulatory units can visualize the development trend of various types of problems over time, and identify the time windows or seasonal patterns of frequent problems. For example, by analyzing the frequency of equipment failures in winter over the years, it can be inferred that low temperatures may be a key factor affecting equipment performance, which in turn guides the adjustment of equipment maintenance strategies in winter.

2.2.2 Text summarization and information extraction

In the face of lengthy and detailed inspection reports, text summarization technology can automatically distill the core content of the report, including the main issues, proposed

corrective measures and implementation status. This not only greatly reduces the reading burden on supervisors, but also ensures timely and accurate decision-making. Information extraction techniques are further refined to accurately capture key entities (e.g., where the problem occurred, who was responsible) and quantitative information (e.g., severity of the problem, cost of rectification) in the report. This highly structured information provides a solid foundation for subsequent data analysis and decision support.

2.2.3 Risk early warning and assessment

The construction of risk warning models is an advanced application that combines textual analysis with predictive analytics. By learning from historical inspection reports, machine learning algorithms are able to identify complex patterns of relationships between problems and risks and predict possible future risk events. For example, if a certain type of problem occurs frequently and the rectification effect is poor, the model may issue an early warning, prompting the supervisory unit to take intervention measures, such as strengthening training, increasing the frequency of inspections, or adjusting the supervisory strategy, so as to prevent the problem from occurring before it is too late.

2.2.4 Intelligent comparison mechanisms

The establishment of an intelligent matching system between standardized inspection items and the report database is an effective way to improve regulatory efficiency. The system automatically identifies problem descriptions in reports through algorithms to match with standardized inspection lists in the database, which not only quickly locates duplicate problems and reduces repetitive work, but also identifies those that are hidden and trending. For example, if the system finds similar equipment at multiple locations with similar failures, it may suggest design flaws or generalized maintenance problems with that model of equipment, providing direction for supervisory units to centralize their resources to solve the problem.

2.2.5 Knowledge map construction

Knowledge graph, as a structured knowledge storage method, can integrate heterogeneous data from multiple sources, such as unsafe event reports, inspection records, and maintenance histories, to form a highly interconnected knowledge network. Through graphical analysis technology, regulatory agencies can deeply explore the causal relationships and dependencies between various types of events, and identify key nodes and vulnerable links in the system. For example, knowledge graphs can reveal the hidden operational malpractice behind an equipment failure, or point out the effectiveness of a particular maintenance measure, providing in-depth insights for accurate supervision. In addition, the recommendation system based on the graph can also propose targeted improvement measures to guide the optimal allocation of regulatory resources.

Text technology plays a crucial role in the analysis of problems found in the inspection, which not only improves the efficiency and quality of supervision, but also lays a solid technical foundation for realizing intelligent supervision and preventive supervision [8]. With the continuous progress of technology and application of in-depth exploration, the future regulatory system will be more intelligent and efficient, and vigorously protect the safe and stable operation of various industries.

3. Building a Regulatory Platform

3.1 Data Collection and Integration

At the initial stage of building a smart regulatory platform, the comprehensiveness and integration of data is its lifeblood. This means not only covering traditional data sources such as unsafe incident reports and inspection reports, but also expanding to emerging information channels such as social media, online forums, and internal communication records, with a view to capturing broader and more detailed regulatory signals. In order to achieve this, the platform needs to deploy an efficient data collection system, which should be equipped with a variety of data acquisition methods, such as automatic crawling, API interface access, user submission, etc., so as to ensure the diversity and real-time nature of data sources.

At the data integration stage, the problem of different data formats and standards needs to be addressed. The platform needs to adopt a unified data model to standardize the data collected, including data format conversion, data standardization and metadata

management. Through the construction of data centers or data warehouses, centralized data storage and management can be achieved, enabling data from different sources and with different structures to "talk" and laying the foundation for subsequent analysis.

3.2 Data Preprocessing and Cleaning

Data preprocessing is an indispensable step before data analysis, and its goal is to improve the quality of data and ensure the accuracy and reliability of analysis results. This process involves a number of links: first, remove invalid data (e.g., blank records, duplicate records), correct erroneous data (e.g., outliers, contradictory information), and deal with missing values by data cleaning, using strategies such as interpolation, deletion, or estimation, to ensure data integrity. For Chinese text data, it is also necessary to perform lexical processing to cut continuous sentences into meaningful word units to provide basic elements for subsequent analysis; at the same time, lexical annotation is performed to distinguish between nouns, verbs, adjectives, etc., which helps to deeply understand the text context.

3.3 Text Analysis Model Construction

Text analytics models are the core components of smart regulation, aiming to mine valuable information from massive text data. Among them, several key technology models are listed below:

1) Text Classification Model: With the help of classical machine learning methods such as SVM and Random Forest, as well as BERT models based on deep learning, the classification of unsafe events and inspection reports can be automated [9]. For example, different types of incidents such as equipment failures, operational errors, environmental factors, etc. can be distinguished, providing a basis for formulating targeted regulatory strategies.

2) Topic models: Utilizing topic models such as LDA (Latent Dirichlet Allocation), potential themes or topics are extracted from a large number of documents, revealing the intrinsic connection of risk factors and helping regulators to grasp industry risk trends and potential problem areas.

3) Named Entity Recognition (NER): By training specialized NER models, key entities such as names of people, organizations, time, and location are extracted from the text, and this information is crucial for tracing the source of the problem and assessing the scope of impact.

4) Sentiment analysis model: quantifying the emotional tendencies (positive, negative or neutral) expressed in texts through sentiment analysis, assessing the severity of the incident and public sentiment, and providing a reference for emergency response and public relations management.

3.4 Risk Management and Early Warning Systems

1) Based on the results of the text analysis, the smart regulatory platform needs to build a set of perfect risk management and early warning mechanism to ensure the foresight and initiative of regulation.

2) Risk assessment: Synthesize the output of text analysis, including classification of issues, distribution of topics, and sentiment tendencies, to conduct a comprehensive risk assessment of the subject of supervision. Through the establishment of a risk scoring system, the risk level of each aspect is quantified, and high-risk areas and urgent problems are identified.

3) Real-time Early Warning: Develop predictive models to monitor the newly collected data in real time, and once an abnormal rise in risk indicators or the emergence of a specific risk pattern is detected, the early warning system will be triggered immediately to notify the relevant departments to take countermeasures in a timely manner.

4) Rectification tracking: A tracking mechanism for the rectification of problems has been established to record information on the progress of rectification, the person responsible, and the expected completion time for each problem, and to verify the effectiveness of rectification through regular return visits and on-site verification. For cases with lagging rectification, key supervision is implemented to ensure that risks are substantially reduced.

3.5 Visualization and Report Generation

The ultimate purpose of a smart regulatory platform is to assist decision-making, so it is particularly important to effectively communicate the analysis results. Through data visualization technology, complex

analysis results are transformed into intuitive charts, dashboards, maps, etc., so that supervisors can quickly understand the distribution characteristics, evolutionary trends and severity of problems. For example, heat maps are used to show areas of high problem incidence, line graphs to track the change of problems over time, and pie charts to analyze the percentage of problem types.

The report generation function should support customization requirements, allowing users to choose the dimension of concern, time frame and presentation format, and automatically generate professional and detailed regulatory reports. These reports should not only contain statistical analysis results, but also insightful recommendations based on data analysis, providing management with a scientific basis for decision-making and promoting the continuous improvement of regulatory effectiveness.

4. Conclusion

By using text technology to analyze unsafe incident reports and inspection findings, the efficiency and accuracy of supervision can be improved and precise supervision can be realized. This paper describes the specific application of text analytics in unsafe incident reports and inspection reports, and discusses the construction of a smart regulatory platform for risk assessment and early warning. In the future, with the continuous progress of technology, text technology will play an increasingly important role in improving aviation safety and realizing precise supervision.

References

[1] Xu Y. The Construction of Ecological Teaching Model of Reading and Appreciation for Ethnic Preparatory Course. Curriculum and Teaching Methodology, 2023, 6(12): 1-5.

[2] Zhang Y, Wei L, Liu X, et al. Nanocellulose/N, O co-doped graphene composite hydrogels for high gravimetric and volumetric performance symmetric supercapacitors. International Journal of Hydrogen Energy, 2022, 47(79): 33827-33838.

[3] Pidgeon N, O'Leary M. Man-made disasters: why technology and organizations (sometimes) fail. Safety science, 2000, 34(1-3): 15-30.

[4] Kao A, Poteet S R. Natural language processing and text mining. Springer Science & Business Media, 2007.

[5] Jia N, Zheng C, Sun W. Speech synthesis of children's reading based on cycle GAN model//Journal of Physics: Conference Series. IOP Publishing, 2020, 1607(1): 012046.

[6] Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research, 2008, 9(2).

[7] Roßbach P. Neural Networks vs. Random Forests–Does it always have to be Deep Learning?. Germany: Frankfurt School of Finance and Management, 2018.

[8] de Souza C S, Leitão C F, Prates R O, et al. Can inspection methods generate valid new knowledge in HCI? The case of semiotic inspection. International Journal of Human-Computer Studies, 2010, 68(1-2): 22-40.

[9] Tixier A J P, Hallowell M R, Rajagopalan B, et al. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction, 2016, 62: 45-56.

