

Original Research Paper

Text Mining of Final Aviation Accident Reports: A Case Study of the Civil Aviation Organization of Iran

Parham Hassan^{1*}  and Hamid Reza Zarghami² 

1- Faculty of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran

2- Faculty of Graduate Studies, Shahid Sattari University of Aerospace Science and Technology, Tehran, Iran

ARTICLE INFO

ABSTRACT

Article History:

Received 03 October 2025

Revised 07 December 2025

Accepted 09 December 2025

Available Online 28 January 2026

Keywords:

Text mining

Air accidents

Machine learning

Topic modeling

Flight safety

In this study, text mining methods and machine learning algorithms were employed to analyze the final reports of Iranian air accidents during the period 2007 to 2025. The dataset consisted of 139 reports from the civil aviation organization database, which were analyzed after textual preprocessing, including the removal of redundant characters, word normalization, stop-word elimination, and stemming. In the first stage, word frequency analysis and topic modeling using the STM algorithm were applied to identify the main themes of the texts. The results indicated that the dominant topics included engine performance, approach and landing processes, aviation training, and helicopter accidents. Subsequently, accident severity was classified into four main categories using KNN, SVM, naive bayes (NB), decision tree (DT), and random forest (RF) algorithms. Model evaluation showed that the random forest (RF) algorithm achieved the highest performance with an accuracy of 63 percent. These findings suggest that the integration of text mining techniques with machine learning algorithms can serve as an effective tool for uncovering hidden patterns in aviation accident reports and supporting safety-related decision-making. Accordingly, it is recommended that future research focus on developing more comprehensive and standardized accident report databases in order to provide a stronger foundation for conducting more complex and precise analyses. Also, using more advanced methods such as deep learning algorithms can further improve classification accuracy.

* Corresponding Author's E-mail: parham.hassan@ut.ac.ir**How to Cite this Article:**

P. Hassan and H. R. Zarghami, "Text mining of final aviation accident reports: A case study of the Civil Aviation Organization of Iran" *Journal of Technology in Aerospace Engineering*, vol. 10, no. 3, pp. 32-42, 2026, (in Persian), <https://doi.org/10.22034/jtae.10.3.3>.

**COPYRIGHTS**

© 2026 by the authors. Published by ARI. This article is an open access article distributed under the terms and conditions of [The Creative Commons Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)



متن کاوی گزارش های نهایی سوانح هوایی: مطالعه موردی سازمان هواپیمایی کشوری ایران

پرهام حسن^{۱*} و حمیدرضا ضرغامی^۲

۱- کارشناسی ارشد، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران
۲- دانشیار، دانشکده تحصیلات تکمیلی، دانشگاه علوم و فنون هوایی شهید ستاری، تهران، ایران

چکیده

در این پژوهش با هدف تحلیل محتوای گزارش های نهایی سوانح هوایی ایران طی بازه ۲۰۰۷ تا ۲۰۲۵، از روش های متن کاوی و الگوریتم های یادگیری ماشین استفاده شده است. داده ها شامل ۱۳۹ گزارش از پایگاه داده سازمان هواپیمایی کشور بوده که پس از پیش پردازش متنی شامل حذف علائم اضافی، یکسان سازی واژگان، حذف ایست واژه ها و ریشه یابی کلمات، مورد تحلیل قرار گرفتند. در بخش نخست، تحلیل فراوانی واژگان و سپس مدل سازی موضوعی (Topic Modeling) به وسیله الگوریتم STM برای شناسایی مضامین اصلی متن ها به کار گرفته شد. نتایج نشان داد که موضوعات شامل عملکرد موتور، فرآیند تقرب و فرود، آموزش هوانوردی و سوانح بالگردی هستند. در ادامه، دسته بندی شدت سوانح در چهار گروه اصلی با بهره گیری از الگوریتم های نزدیک ترین همسایگی، ماشین بردار پشتیبان، بیز ساده، درخت تصمیم و جنگل تصادفی انجام گرفت. ارزیابی مدل ها نشان داد که الگوریتم جنگل تصادفی با دقت ۶۹ درصد عملکرد بهتری نسبت به سایر مدل ها داشته است. نتایج این مطالعه بیانگر آن است که به کارگیری روش های متن کاوی در کنار الگوریتم های یادگیری ماشین می تواند ابزاری کارآمد برای شناسایی الگوهای پنهان در گزارش های سوانح هوایی و پشتیبانی از تصمیم گیری های ایمنی باشد. بر این اساس، پیشنهاد می شود در پژوهش های آینده بانک های داده ای جامع تر و استانداردتر از گزارش های سوانح فراهم گردد تا بستری مناسب برای انجام تحلیل های پیچیده تر و دقیق تر فراهم شود. همچنین بهره گیری از روش های پیشرفته تر مانند الگوریتم های یادگیری عمیق نیز می تواند دقت طبقه بندی را بهبود بیشتری دهد.

اطلاعات مقاله

تاریخچه مقاله:

دریافت ۱۱ مهر ۱۴۰۴
بازنگری ۱۶ آذر ۱۴۰۴
پذیرش ۱۸ آذر ۱۴۰۴
اولین انتشار ۸ بهمن ۱۴۰۴

واژه های کلیدی:

متن کاوی
سوانح هوایی
یادگیری ماشین
مدل سازی موضوعی
ایمنی پرواز

پست الکترونیکی نویسنده مسئول: parham.hassan@ut.ac.ir

How to Cite this Article:

P. Hassan and H. R. Zarghami, "Text mining of final aviation accident reports: A case study of the Civil Aviation Organization of Iran" *Journal of Technology in Aerospace Engineering*, vol. 10, no. 3, pp. 32-42, 2026, (in Persian), <https://doi.org/1022034/jtae.10.3.3>.



COPYRIGHTS

© 2026 by the authors. Published by ARI. This article is an open access article distributed under the terms and conditions of [The Creative Commons Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)



۱. مقدمه

ایمنی پرواز یکی از مهم‌ترین دغدغه‌های صنعت هوانوردی در سطح ملی و بین‌المللی است. بررسی سوانح و حوادث هوایی نه تنها به شناسایی علل و عوامل مؤثر در وقوع آن‌ها کمک می‌کند، بلکه می‌تواند زمینه‌ساز تدوین راه‌کارهایی برای ارتقای ایمنی و کاهش ریسک در عملیات پروازی باشد. سازمان‌های هوانوردی در کشورهای مختلف همواره تلاش کرده‌اند با گردآوری و انتشار گزارش‌های نهایی سوانح، بستری برای تحلیل و بهره‌برداری پژوهشگران و متخصصان این حوزه فراهم آورند.

در ایران نیز سازمان هواپیمایی کشوری گزارش‌های متعددی از سوانح هوایی را منتشر کرده است که شامل جزئیات فنی، عملیاتی و انسانی مرتبط با حوادث است. این گزارش‌ها حجم بالایی از داده‌های متنی را در بر می‌گیرند که تحلیل دستی آن‌ها زمان‌بر و پرهزینه خواهد بود. در چنین شرایطی، استفاده از روش‌های نوین تحلیل داده و به‌ویژه متن‌کاوی (Text Mining) می‌تواند به کشف الگوها، موضوعات پنهان و روابط میان متغیرها کمک شایانی کند.

متن‌کاوی و پردازش زبان طبیعی (NLP) در سال‌های اخیر در حوزه‌های مختلفی همچون علوم اجتماعی، پزشکی و مدیریت دانش به‌کار گرفته شده‌اند. در حوزه هوانوردی نیز این رویکردها می‌توانند ابزاری قدرتمند برای تحلیل گزارش‌های ایمنی و سوانح باشند. ترکیب متن‌کاوی با الگوریتم‌های یادگیری ماشین این امکان را فراهم می‌سازد که گزارش‌های سوانح هوایی به‌صورت خودکار دسته‌بندی شده و موضوعات کلیدی آن‌ها شناسایی شوند. این امر نه تنها موجب تسریع فرآیند تحلیل می‌شود، بلکه دقت تصمیم‌گیری‌های مبتنی بر داده را نیز افزایش می‌دهد.

هدف پژوهش حاضر، تحلیل گزارش‌های نهایی سوانح هوایی ایران طی بازه زمانی ۲۰۰۷ تا ۲۰۲۵ با استفاده از روش‌های متن‌کاوی و یادگیری ماشین است. در این مطالعه، ابتدا داده‌ها پیش‌پردازش و پاک‌سازی شدند و سپس با بهره‌گیری از روش‌های تحلیل فراوانی واژگان و مدل‌سازی موضوعی، مضامین اصلی متن‌ها استخراج گردید. در ادامه، دسته‌بندی شدت سوانح با الگوریتم‌های مختلف یادگیری ماشین انجام و نتایج مدل‌ها مقایسه شد. نوآوری این پژوهش در به‌کارگیری ترکیبی از روش‌های متن‌کاوی و یادگیری ماشین برای تحلیل جامع گزارش‌های سوانح هوایی داخلی است که می‌تواند به سازمان‌های هوانوردی در بهبود سیاست‌های ایمنی و پیشگیری از سوانح کمک کند.

۲. پیشینه پژوهش

مطالعه و تحلیل سوانح هوایی از دیرباز یکی از موضوعات کلیدی در

حوزه ایمنی پرواز بوده است. در سطح بین‌المللی، پژوهش‌های متعددی به بررسی گزارش‌های سوانح و حوادث هوایی پرداخته‌اند. از جمله این روش‌ها به پردازش زبان طبیعی (NLP) و متن‌کاوی برای استخراج دانش از گزارش‌های حادثه و ایمنی هوانوردی می‌توان اشاره کرد. یکی از این پژوهش‌ها، مروری نظام‌مند بر کاربردهای NLP در ایمنی هوانوردی نشان می‌دهد که مطالعات مروری و چارچوب‌های کلی، نقشه روش‌شناختی این حوزه را ترسیم کرده و رویکردهایی مانند خوشه‌بندی، مدل‌سازی موضوع، تشخیص موجودیت‌های اسمی و مدل‌های زبانی پیش‌آموزش‌داده را به‌عنوان ابزارهای اصلی معرفی کرده‌اند [۱]. در جهت عملی، پژوهشی مجموعه‌ای از گزارش‌های سوانح به زبان چینی را برای ساخت مدل‌ها ترنسفورمر، تجزیه و تحلیل خوشه‌بندی و مدل‌سازی شبکه و ویژگی مورد استفاده قرار داده و نشان داده است که پیش‌آموزش مدل‌های ترنسفورمر روی داده‌های ایمنی هوانوردی باعث ساده‌سازی متون اصلی طولانی و تکراری می‌شود [۲].

در پژوهشی دیگر که متن‌کاوی را با تحلیل شبکه ترکیب کرده است نشان داده که با ایجاد گراف هم‌رخدادی (co-occurrence) بین مفاهیم می‌توان خوشه‌ها و الگوهای هم‌عوامل را بهتر یافت و در نتیجه روابط علی یا هم‌ترازی عوامل انسانی/فنی را روشن‌تر کرد؛ این روش به‌ویژه در شناخت عوامل چندگانه در گزارش‌های طولانی و چندوجهی مفید است [۳]. در زمینه پیش‌بینی و طبقه‌بندی خودکار، مطالعه‌ای از ترکیب ویژگی‌های استخراج‌شده از متن (بردارهای TF-IDF یا embeddings) با الگوریتم‌های طبقه‌بندی و همچنین شبکه‌های عصبی متوالی (LSTM/BLSTM) گزارش کرده است که برای برچسب‌گذاری شدت حادثه یا نوع خطا قابل‌استفاده است [۴]. پژوهش‌های متعددی در حوزه دسته‌بندی گزارشات سوانح هوایی با بکارگیری تکنیک‌های یادگیری ماشین و متن‌کاوی توسط پژوهشگران انجام شده است [۲، ۴-۶]. در پژوهشی از مدل درخت تصمیم تقویت یافته شدید (XGBoost) استفاده شده و نشان داده شده که دقت میانگین ۰/۸۵ برای میانگین نمره F1 به‌دست آمده است [۷]، همچنین در پژوهشی دیگر محققان با بکارگیری الگوریتم جنگل تصادفی (Random Forest) جهت دسته‌بندی خودکار سوانح هوایی استفاده کردند و نشان دادند که این مدل قادر به دسته‌بندی بیش از ۴۵۰۰۰ گزارش با دقت بین ۸۰ تا ۹۳ درصد است [۸]. مطالعه‌ای دیگر کاربرد هوش مصنوعی در دسته‌بندی گزارشات سوانح هوایی را بررسی کرده و نشان داده که هوش مصنوعی می‌تواند دانش‌های جدیدی از سوانح را برای سازمان‌های یادگیرنده ایجاد کند [۹]. پژوهشی دیگر به بررسی مدل‌های گروهی در حوزه یادگیری ماشین برای پیش‌بینی ریسک در سوانح هوایی پرداخته و نشان داده ترکیب مدل ماشین بردار پشتیبان و یادگیری عمیق قادر به تشخیص سطح ریسک در سوانح غیرعادی است

۳-۱ داده‌ها و منابع اطلاعاتی

داده‌های مورد استفاده شامل ۱۳۹ گزارش نهایی سوانح هوایی منتشر شده در پایگاه‌های رسمی سازمان هواپیمایی کشوری [۱۶] و گروه بررسی سوانح است. تمام گزارش‌های موجود به‌جز مواردی که فایل گزارش موجود نبوده مورد استفاده قرار گرفته است. این گزارش‌ها بازه زمانی ۲۰۰۷ تا ۲۰۲۵ را پوشش می‌دهند و از نظر شدت در چهار دسته اصلی طبقه‌بندی شده‌اند که در جدول ۱ به همراه فراوانی قابل مشاهده است. تعریف عملیاتی شدت سوانح بر اساس پیوست ۱۳ سازمان بین‌المللی هوانوردی غیرنظامی (ICAO) به شرح زیر است [۱۷]:

سانحه (Accident): به رویدادی گفته می‌شود که در جریان عملیات یک هواپیما رخ دهد و منجر به جراحت مرگبار یا جدی افراد، آسیب جدی و مؤثر به ساختار یا عملکرد هواپیما، یا مفقود شدن کامل هواپیما شود. این تعریف برای هواپیماهای سرنشین‌دار از زمان ورود افراد به قصد پرواز تا خروج همه آن‌ها، و برای هواپیماهای بدون سرنشین از زمان آماده‌سازی برای پرواز تا توقف کامل ادامه دارد.

حادثه (Incident): رویدادی غیر از سانحه که در ارتباط با عملیات یک هواپیما اتفاق می‌افتد و بر ایمنی عملیات تأثیر می‌گذارد یا می‌تواند تأثیر بگذارد.

حادثه جدی (Serious Incident): رویدادی است که احتمال تبدیل شدن آن به سانحه بسیار بالا بوده اما در نهایت به آسیب جدی یا تلفات منتهی نشده است. بازه زمانی وقوع آن همانند سانحه است. تفاوت اصلی با سانحه این است که پیامدهای شدید در آن رخ نداده، اما خطر وقوعشان کاملاً وجود داشته است.

نزدیک پروازی در رده حادثه جدی (Airprox-Serious Incident): رویداد تقارب خطرناک میان هواپیماها که جدایی ایمن رعایت نشده و احتمال برخورد بسیار بالا بوده، اما رویداد در نهایت به سانحه منجر نشده است.

جدول ۱ - فراوانی و درصد فراوانی شدت سوانح هوایی ۱۳۹ گزارش سانحه هوایی از پایگاه داده سازمان هواپیمایی کشور.

Table 1. Frequency and severity percentage of air accidents of 139 air accident reports from civil aviation organization database.

	Accident Severity	Frequency	Frequency percentage
1	Accident	69	49.6
2	Incident	35	25.1
3	Serious Incident	29	20.8
4	Airprox-Serious Incident	6	0.04

[۱۰]. همچنین در حوزه پیشبینی جراحات و مرگ و میر در سوانح هوایی پژوهشی با بکارگیری یادگیری ماشین و استفاده از مدل‌های مختلف نشان داده که شبکه‌های عصبی مصنوعی دقیق‌ترین پیشبینی را برای سوانح شامل مرگ و میر یا سوانح با شدت بالا دارد [۱۱]. همچنین نتایج این پژوهش نشان داده است که تکنیک‌های یادگیری ماشین ممکن است اطلاعات بهتری از طریق روش‌های تحلیل آماری سنتی ارائه دهند.

در حوزه شناسایی ریسک‌های ایمنی هوانوردی، برخی پژوهش‌ها چارچوب‌های عملیاتی مبتنی بر متن‌کاوی طراحی کرده‌اند. در این رویکردها ابتدا گزارش‌های حادثه گردآوری و پیش‌پردازش شده‌اند تا عبارات مرتبط با عوامل خطر (hazards) و شرایط ناایمن شناسایی شوند. سپس با استفاده از روش‌های مدل‌سازی موضوع و استخراج کلید واژه‌ها، مجموعه‌ای از ریسک‌های بالقوه استخراج شده است [۱۲، ۱۳].

در ایران اما تاکنون پژوهش‌های محدودی در زمینه تحلیل متنی گزارش‌های نهایی سوانح هوایی انجام شده است. مطالعات داخلی عمدتاً بر داده‌های آماری سوانح و استفاده از روش‌های سنتی تحلیل داده متمرکز بوده‌اند [۱۴]. برای مثال، پژوهشی [۱۵] با بهره‌گیری از نرم‌افزارهایی مانند SPSS Modeler و SPSS، داده‌های سوانح هوانوردی را تحلیل کرده و با استفاده از الگوریتم‌هایی همچون درخت تصمیم، الگوهای حوادث را بر اساس ویژگی‌هایی مانند زمان، مکان، نوع هواپیما و تجربه خلبان شناسایی کرده‌اند. این مطالعات محدود به تحلیل‌های آماری یا بررسی‌های موردی خاص هستند و کمتر از روش‌های نوین متن‌کاوی و یادگیری ماشین بهره برده‌اند. این خلأ پژوهشی ضرورت انجام تحقیق حاضر را روشن می‌سازد؛ چراکه ترکیب تکنیک‌های متن‌کاوی با الگوریتم‌های یادگیری ماشین می‌تواند به کشف بینش‌های جدید از گزارش‌های نهایی سوانح هوایی ایران منجر شود و مبنای تصمیم‌گیری‌های مبتنی بر داده در حوزه ایمنی پرواز را فراهم کند.

۳. روش‌شناسی پژوهش

روش‌شناسی پژوهش حاضر بر مبنای رویکرد متن‌کاوی (Text Mining) و یادگیری ماشین (Machine Learning) طراحی شده است. این چارچوب با هدف استخراج الگوهای پنهان از متون گزارش‌های نهایی سوانح هوایی و دسته‌بندی شدت سوانح در چند گام متوالی پیاده‌سازی شد:

- گردآوری و آماده‌سازی داده‌ها
- پیش‌پردازش و پاک‌سازی متون
- تحلیل اکتشافی و مدل‌سازی موضوعی
- دسته‌بندی شدت سوانح با الگوریتم‌های یادگیری ماشین
- ارزیابی و تحلیل نتایج

۳-۲ پیش پردازش متون

از آنجا که گزارش‌ها به صورت فایل‌های PDF منتشر می‌شوند لازم است تا ابتدا فایل‌های PDF به کمک نرم‌افزار Python و پکیج pdfplumber به فایل‌های TXT تبدیل شود. سپس مقدمات پیش‌پردازش متون از طریق زبان برنامه‌نویسی R برای آماده‌سازی داده‌ها انجام شود.

مراحل پیش پردازش داده‌ها شامل بخش‌های زیر است:

- استخراج متن از PDF: با استفاده از کتابخانه پایتون pdfplumber (v 0.11.04) و تبدیل آن‌ها به فایل‌های TXT از طریق تابع `extract_text`، متن هر صفحه را استخراج کرده و خروجی نهایی در قالب فایل‌های TXT ذخیره شد.
- نرمال‌سازی کاراکترها: این بخش شامل کوچک‌سازی حروف انگلیسی، حذف کاراکترهای زائد و فاصله‌های تکراری است که از توابع `tolower` و `str_replace_all` پکیج `stringr` (v 1.5.1) استفاده شده است.
- حذف علائم نگارشی: شامل حذف نقطه، ویرگول، پرانتز و نشانه‌های کنترلی می‌شود. در این بخش از توابع `removePunctuation` و `str_replace_all` و `removeNumbers` در پکیج‌های `stringr` (v 1.5.1) و `tm` (v 0.7.16) استفاده شده است.
- حذف واژه‌های توقف (Stop Words): شامل واژگانی فاقد ارزش معنایی است. برای متون انگلیسی از توابع `stopwords` در پکیج `tm` و برای متون فارسی شامل کلمات متداول مانند «از»، «با»، «به» و «در» و غیره است.
- قطعه‌بندی متن (Tokenization): شامل شکسته شدن متن و جملات به واحدهای واژه‌ای است. که از روش فاصله بین کلمات از طریق تابع `str_split` در پکیج `stringr` انجام شده است.
- ریشه‌یابی (lemmatization/stemming): بازگرداندن واژه‌ها به شکل پایه که از طریق تابع `persian_stemmer` در بسته نرم‌افزاری `PersianStemmer` (v 1.0) برای فارسی و `stemDocument` در پکیج `tm` برای انگلیسی انجام شده است.
- برداری‌سازی (Vectorization): تبدیل متون به بردارهای عددی با روش TF-IDF به منظور استفاده در مدل‌های آماری و یادگیری ماشین.
- این مجموعه اقدامات، داده‌ها را از حالت خام به شکلی استاندارد و قابل تحلیل تبدیل کرد.

۳-۳ تحلیل فراوانی واژگان

این روش به شناسایی و شمارش تکرار واژگان در مجموعه متنی

می‌پردازد و امکان بررسی الگوهای غالب در اسناد را فراهم می‌آورد. تحلیل فراوانی واژگان کمک می‌کند تا واژگان کلیدی و مفاهیم غالب در داده‌ها شناسایی شوند و دید اولیه‌ای نسبت به موضوعات و گرایش‌های اصلی در متن ارائه شود. این مرحله مقدماتی، پایه‌ای برای انجام تحلیل‌های پیچیده‌تر و استخراج موضوعات فراهم می‌کند.

۴-۴ مدل‌سازی موضوعی

مدل‌سازی موضوعی (Topic Modeling) یک تکنیک یادگیری ماشین است که برای کشف الگوهای پنهان در مجموعه‌های متنی طراحی شده است و دیدگاه‌هایی درباره محتوای اصلی، ساختارهای معنایی پنهان و موضوعات کلیدی ارائه می‌دهد [۲]. این روش، داده‌های متنی را به عنوان مجموعه‌ای از موضوع‌ها و توزیع واژگان مرتبط با هر موضوع در نظر می‌گیرد. برخلاف الگوریتم‌های سنتی طبقه‌بندی، مدل‌سازی موضوعی امکان وجود چندین موضوع برای هر سند را فراهم می‌کند و بر کشف الگوهای زیرین تمرکز دارد، برخلاف مدل‌های بانظر که بر اختصاص دادن دسته‌های از پیش تعیین شده متمرکز هستند. به دلیل ماهیت یادگیری بدون ناظر این روش، مدل‌سازی موضوعی نیازی به برچسب‌ها یا دسته‌بندی‌های از پیش تعریف شده ندارد.

۴-۵ مدل موضوعی ساختاری

مدل STM یک مدل مبتنی بر احتمالات است که به منظور استخراج موضوعات پنهان از مجموعه‌ای از اسناد طراحی شده است. این مدل قادر است تا توزیع موضوعات در اسناد را شناسایی کند. یکی از ویژگی‌های کلیدی STM، جایگزینی توزیع پیشین دریکله با یک توزیع نرمال لجستیک است که این امر اجازه می‌دهد تا وابستگی‌ها و همبستگی‌های بین موضوعات به شکلی دقیق‌تر مدل‌سازی شود. در این مدل، پارامترهای مربوط به توزیع موضوعات و کلمات، تحت تأثیر متغیرهای ساختاری قرار می‌گیرند که این موضوع باعث بهبود انعطاف‌پذیری و دقت مدل در استخراج موضوعات مرتبط با شرایط مختلف اسناد می‌شود.

۴-۶ دسته بندی متن

دسته‌بندی متن (Text Classification) یکی از مهم‌ترین شاخه‌های پردازش زبان طبیعی (NLP) و یادگیری ماشین است که هدف آن اختصاص دادن اسناد متنی به چند برچسب یا دسته از پیش تعیین شده می‌باشد. در واقع، این روش برخلاف مدل‌سازی موضوعی که یک رویکرد بدون ناظر است، در چارچوب یادگیری نظارت‌شده (Supervised Learning) قرار می‌گیرد، زیرا مجموعه‌ای از اسناد برچسب‌گذاری شده برای آموزش مدل مورد نیاز است.

در فرآیند دسته‌بندی متن، مدل می‌آموزد که با توجه به الگوهای

به‌طور شهودی، اگر داده‌ها به‌صورت خطی قابل جداسازی باشند، SVM یک خط (در دو بعد)، یک صفحه (در سه بعد) یا به‌طور کلی یک ابرصفحه در فضای چندبعدی می‌سازد. اگر داده‌ها به‌صورت خطی قابل جداسازی نباشند، SVM با استفاده از توابع هسته‌ای (Functions Kernel) داده‌ها را به فضای ابعاد بالاتر نگاشت می‌کند تا در آن فضا بتوان مرز خطی بهتری برای جداسازی پیدا کند.

۳-۶-۳ بیز ساده

الگوریتم بیز ساده بر پایه قضیه بیز در معادله (۲) و فرض استقلال شرطی بین ویژگی‌ها عمل می‌کند. این الگوریتم به‌دلیل سادگی محاسبات سرعت بیشتری نسبت به سایر مدل‌های یادگیری ماشین دارد. ایده اصلی این است که احتمال تعلق یک متن به یک کلاس خاص، بر اساس احتمال وقوع ویژگی‌های آن (مثلاً کلمات موجود در متن) محاسبه می‌شود. البته که فرض استقلال کامل بین ویژگی‌ها در عمل همیشه الزاماً برقرار نیست. احتمال قرار گرفتن در کلاس C براساس معادله (۳) با فرض استقلال به‌دست می‌آید.

$$p(Y = C|X) = \frac{P(X|Y = C)P(C)}{P(X)} \quad (2)$$

$$P(Y = C|x_1, x_2, \dots, x_p) = \frac{P(C) \prod_{i=1}^p P(x_i|C)}{P(x_1, x_2, \dots, x_p)} \quad (3)$$

۳-۶-۴ درخت تصمیم

درخت تصمیم یک روش شهودی و قابل تفسیر برای طبقه‌بندی داده‌ها است که در مسئله رگرسیون و طبقه‌بندی مورد استفاده قرار می‌گیرد. در این روش، داده‌ها به‌صورت سلسله مراتبی و بر اساس ویژگی‌ها تقسیم می‌شوند تا در نهایت به یک گره تصمیم برسیم. هر گره داخلی درخت بیانگر یک ویژگی، هر شاخه یک شرط روی آن ویژگی و هر برگ نشان دهنده برچسب کلاس است.

درخت تصمیم به‌دلیل ساختار ساده و قابلیت تفسیر بالا، یکی از الگوریتم‌های پرکاربرد در یادگیری ماشین است. برای ساخت درخت، معمولاً از معیارهایی مانند انترپوی در معادله (۴) و شاخص جینی در معادله (۵) برای انتخاب بهترین ویژگی در هر گره استفاده می‌شود.

$$E(S) = - \sum_{i=1}^c p_i \log_2(p_i) \quad (4)$$

$$G(S) = 1 - \sum_{i=1}^n p_i^2 \quad (5)$$

واژگانی، آماری و معنایی در داده‌های آموزشی، اسناد جدید و ناشناخته را در دسته‌های درست قرار دهد. این تکنیک در حوزه‌های مختلفی مانند تحلیل احساسات نظرات کاربران در شبکه‌های اجتماعی با برچسب‌های مثبت، منفی و خنثی یا فیلتر کردن هرزنامه‌ها (Spam Filtering) به‌ویژه در تشخیص ایمیل‌ها استفاده می‌شود. همچنین در دسته‌بندی خودکار گزارش‌های حوادث به‌کار گرفته می‌شود. در ادامه به بررسی متداول‌ترین مدل‌های مورد استفاده در بخش یادگیری با ناظر می‌پردازیم.

۳-۶-۱ نزدیک‌ترین همسایگی

الگوریتم نزدیک‌ترین همسایگی (Nearest Neighbour) یکی از ساده‌ترین و در عین حال کاربردی‌ترین روش‌های یادگیری با ناظر است. در این روش فرض بر این است که داده‌هایی که در فضای ویژگی‌ها به هم نزدیک‌ترند، برچسب مشابهی دارند. برای طبقه‌بندی یک نمونه جدید، ابتدا فاصله آن تا تمام داده‌های آموزشی محاسبه می‌شود و سپس بر اساس برچسب غالب میان نزدیک‌ترین همسایه‌ها، کلاس نهایی انتخاب می‌شود.

این روش به‌دلیل سادگی و عدم نیاز به مرحله آموزش پیچیده، در بسیاری از مسائل طبقه‌بندی مورد استفاده قرار می‌گیرد. از سوی دیگر، یکی از معایب KNN این است که با افزایش حجم داده‌ها، محاسبات فاصله زمان‌بر می‌شود و در داده‌های با ابعاد بالا (مانند متن که شامل هزاران ویژگی می‌شود) ممکن است دقت کاهش یابد. همچنین در این روش از فاصله‌های مختلفی می‌توان استفاده کرد که یکی از متداول‌ترین آن‌ها فاصله اقلیدسی است. فرض کنید بردار ویژگی‌های $x = (x_1, x_2, \dots, x_p)$ و $y = (y_1, y_2, \dots, y_p)$ باشد. در این صورت فاصله دو بردار بر اساس معادله (۱) محاسبه می‌شود.

$$d(x, y) = \sum_{i=1}^p \sqrt{(x_i - y_i)^2} \quad (1)$$

۳-۶-۲ ماشین بردار پشتیبان

ماشین بردار پشتیبان (SVM) یکی از الگوریتم‌های قدرتمند یادگیری با ناظر است که در مسائل طبقه‌بندی بسیار متداول است. ایده اصلی SVM این است که یک ابرصفحه (Hyperplane) پیدا کند که داده‌های متعلق به کلاس‌های مختلف را از هم جدا کند. بهترین ابرصفحه، آن است که بیشترین فاصله (Margin) را از نزدیک‌ترین نقاط هر کلاس داشته باشد. این نقاط نزدیک به مرز، «بردارهای پشتیبان» نامیده می‌شوند و نقش کلیدی در تعیین مکان مرز تصمیم‌گیری دارند.

جنگل تصادفی یک روش ترکیبی (Ensemble) مبتنی بر مجموعه‌ای از درخت‌های تصمیم است. در این الگوریتم، هر درخت با نمونه‌گیری تصادفی از داده‌ها و انتخاب تصادفی از ویژگی‌ها ساخته می‌شود. سپس پیش‌بینی نهایی از طریق رأی‌گیری اکثریت میان درخت‌ها تعیین می‌شود.

مزیت اصلی جنگل تصادفی نسبت به درخت تصمیم منفرد، کاهش بیش‌برازش (Overfitting) و افزایش دقت است. همچنین، این الگوریتم می‌تواند بر اهمیت نسبی ویژگی‌ها نیز اطلاعاتی ارائه دهد که در تحلیل متن می‌تواند بسیار ارزشمند باشد.

شکل ۱- ابر کلمات استخراج شده از گزارش های سوانح هوایی (متون فارسی).

*با توجه به اظهار نویسنده مبنی بر استخراج اطلاعات از متون فارسی امکان درج معادل انگلیسی آن‌ها در شکل میسر نمی‌باشد.



Fig. 2. Word cloud extracted from aviation accident reports (English texts)

جهت شناسایی مضامین اصلی موجود در گزارش‌های نهایی سوانح هوایی، از مدل STM استفاده شده است. خروجی این مدل نشان می‌دهد که گزارش‌ها قابل تجزیه به هشت موضوع متمایز هستند که هر یک بر جنبه‌ای از عوامل مؤثر در بروز سوانح تمرکز دارند. واژگان کلیدی شامل احتمال بیشینه (Highest Probability) و واژگان متمایزکننده (FREX) استخراج شده از هر موضوع در جدول ۲ ارائه شده است همچنین تفسیر و خلاصه موضوعات نیز در جدول ۳ ذکر شده است. مقادیر مختلف برای پارامتر تعداد موضوعات آزموده شد تا ساختار پنهان متن‌ها انسجام مناسب داشته باشد. خروجی‌های بررسی شده نشان داد که مقدار هشت موضوع بهترین توازن را میان انسجام درونی

برای سنجش دقت مدل‌های یادگیری ماشین در دسته‌بندی شدت سوانح هوایی، از مجموعه‌ای از معیارهای ارزیابی متداول در حوزه طبقه‌بندی متون استفاده شده است. این معیارها شامل دقت (Accuracy)، در معادله (۷)، کاپای کوهن (Cohen's Kappa) در معادله (۸)، امتیاز F1 ماکرو (F1-Macro) در معادله (۱۰) بر اساس F1 در معادله (۹) هستند. ارزیابی مدل با تقسیم داده‌های آموزشی (Train) و آزمایشی (Test) با نسبت‌های ۶۰ و ۴۰ درصد تقسیم شده و همچنین تنظیم پارامترها نیز با آزمون و خطا انجام شده است.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (\lambda)$$

$$F1 - Macro = \frac{1}{C} \sum_{i=1}^C F1_i \quad (10)$$

۴-۱ تحلیل فراوانی، واژه‌ها

نتایج در این بخش نشان می‌دهد که در متون فارسی، واژه‌هایی همچون «پرواز»، «فروودگاه»، «خلبان» و «سانحه» بیشترین تکرار را دارند، در حالی که در متون انگلیسی واژه‌های «aircraft»، «flight»، «pilot» و «engine» پرتکرارتر هستند. شکل ۱ و شکل ۲ به ترتیب ابر کلمات متون فارسی و متون انگلیسی را نمایش می‌دهند.

جدول ۳- نتایج مدل‌سازی موضوعی با الگوریتم STM به‌همراه خلاصه موضوع و درصد فراوانی موضوعات استخراج‌شده از گزارش‌های سوانح هوایی.

Table 3. Topic modeling results with STM algorithm along with topic summary and frequency percentage of topics extracted from aviation accident reports.

#	Main title	Summary	Frequency percentage
1	تحلیل سوانح نظامی	گزارش‌های تحلیلی و بررسی‌های سوانح مرتبط با موضوع‌های نظامی است.	2.74
2	حوادث بالگرد و هلیکوپتر	این موضوع به حوادث و بررسی‌های مرتبط با هلیکوپترها و بالگردها مربوط است.	10.77
3	عملکرد پرواز و گزارش‌های انگلیسی	گزارش‌های انگلیسی که درباره عملکرد موتور، فرود، خلبان، خدمه و کنترل‌های فرودگاهی است.	28.47
4	تقرب، فرود و کنترل باند	این موضوع به مراحل تقرب و فرود در باندهای پروازی، هماهنگی‌های رادیویی و نظارت‌های فرودگاهی مربوط می‌شود.	13.11
5	مشکلات فنی هواپیما	این موضوع به مسائل فنی پرواز مانند موتور، سیستم‌ها، شتاب، ارتفاع و تغییرات عملکرد هواپیما در حین پرواز اشاره دارد.	4.00
6	پروازهای آموزشی و محلی	این موضوع به پروازهای آموزشی یا داخلی (نظیر پروازهای در شهرهای مختلف مثل نوشهر، گلبهار و غیره) می‌پردازد.	6.86
7	عملیات و تصمیم‌گیری پروازی	این موضوع به استفاده عملیاتی از هواپیماها توسط شرکت‌ها، کنترلرها و خلبانان، همچنین تصمیم‌گیری در شرایط باند و عملیات پروازی مربوط است.	18.62
8	آموزش هوانوردی و سازماندهی	این موضوع به حوزه آموزش هوانوردی (دانشجویان، معلمان، گواهینامه‌های خلبانی) و جنبه‌های فنی و سازمانی مربوط به آن‌ها اشاره دارد.	15.42

*با توجه به اظهار نویسنده مبنی بر استخراج اطلاعات از متون فارسی امکان درج معادل انگلیسی آن‌ها در جدول میسر نمی‌باشد.

۴-۳ دسته‌بندی متون

برای پیش‌بینی و دسته‌بندی شدت سوانح هوایی از روی محتوای گزارش‌های نهایی، از الگوریتم یادگیری ماشین شامل نزدیک‌ترین

واژگان، تفکیک‌پذیری مفهومی و قابلیت تفسیر فراهم می‌کند. در مقادیر کمتر از هشت، موضوعات ایجاد شده بیش از حد گسترده بودند و چندین مضمون ناهمگون در یک موضوع ادغام می‌شدند؛ در نتیجه، مرزبندی معنایی بین بخش‌های مختلف گزارش‌ها کاهش می‌یافت. در مقابل، مقادیر بالاتر از هشت منجر به تولید موضوعات پراکنده و کم‌معنا می‌شد که از نظر محتوایی وزن و ثبات کافی نداشتند و بخشی از آن‌ها تنها شامل واژگان وابسته به نویز یا محتوای کم‌تکرار بودند. این نتایج نشان داد که تعداد موضوعات در مقدار هشت، بیشترین انطباق را با لایه‌های معنایی واقعی موجود در گزارش‌های سوانح دارد و دسته‌بندی حاصل نیز از نظر عملیاتی و فنی قابل تفسیرتر است. بر این اساس، مقدار ۸ موضوع به‌عنوان گزینه بهینه انتخاب شد.

جدول ۲- موضوعات استخراج‌شده و واژگان شاخص بر اساس معیارهای احتمال بیشینه و واژگان متمایز کننده با استفاده از مدل STM.

Table 2. Topics extracted using the STM model and their key terms based on highest probability and frex criterias.

#	Main title	Highest probability	FREX
1	تحلیل سوانح نظامی	نظامی، هوایی، شکل، سوانح، دفتر	نظامی، شکل، هواپیمایی، اسفند، هوایی
2	حوادث بالگرد و هلیکوپتر	هلیکوپتر، بالگرد، هلیکوپتری، کارخانه، اداره	هلیکوپتر، بالگرد، شرکت، اداره، پرواز
3	عملکرد پرواز و گزارش‌های انگلیسی	Engin, Airport, Accid, Pilot, Crew	Engin, Airport, Accid, Pilot, Oper
4	تقرب، فرود و کنترل باند	پروازی، باند، فرود، تقرب، صدور	صدور، تقرب، نالملی، محلی، رویداد
5	مشکلات فنی هواپیما	هواپیما، موتور، سیستم، میزان، ارتفاع	سیستم، میزان، شتاب، موتور، افزایش
6	پروازهای آموزشی و محلی	هواپیما، اطلاعات، موتور، گلبهار، کشوری	هواپیمای، گلبهار، کشوری، پیام، نوشهر
7	عملیات و تصمیم‌گیری پروازی	هواپیما، باند، هواپیمایی، اطلاعات، پیش‌رانه	هواپیمای، پرنده، کنترلر، تاکسی، تصمیم
8	آموزش هوانوردی و سازماندهی	دانشجو، معلم، خلبانی، فنی، آموزشی	فنی، خلبانی، آموزشی، گواهینامه، دانشجو

*با توجه به اظهار نویسنده مبنی بر استخراج اطلاعات از متون فارسی امکان درج معادل انگلیسی آن‌ها در جدول میسر نمی‌باشد.

جدول ۴- مقایسه عملکرد مدل‌های یادگیری ماشین در دسته‌بندی شدت سوانح بر اساس معیارهای ارزیابی شامل دقت (بازه اطمینان ۹۵ درصد)، کاپا و F1 ماکرو.

Table 4. Comparison of machine learning models' performance in classifying accident severity based on evaluation metrics including Accuracy (95% confidence interval), Kappa, and Macro F1.

Model	Accuray (CI 95%)	Kappa	F1_Macro
KNN	0.581 (0.441, 0.715)	0.195	0.302
SVM	0.545 (0.405, 0.680)	0.111	0.246
NB	0.490 (0.353, 0.629)	-0.002	0.198
DT	0.418 (0.286, 0.558)	0.045	0.265
RF	0.690 (0.551, 0.808)	0.448	0.467

جدول ۵- مقایسه عملکرد مدل‌های یادگیری ماشین در طبقه‌بندی هر کلاس شدت سوانح هوایی بر اساس معیارهای درستی و F1.

Table 5. Comparison of machine learning models' performance in classifying each air accident severity class based on Precision and F1 metrics.

Class	Model	Precision	F1
Accident	KNN	0.571	0.727
	SVM	0.549	0.708
	NB	0.500	0.650
	DT	0.531	0.566
	RF	0.682	0.811
Incident	KNN	0.600	0.315
	SVM	0.500	0.125
	NB	0	0.00
	DT	0.363	0.320
	RF	0.555	0.434
Serious incident	KNN	1.000	0.166
	SVM	0.500	0.153
	NB	0.333	0.142
	DT	0.166	0.173
	RF	1.000	0.625
Airprox-serious incident	KNN	0	0
	SVM	0	0
	NB	0	0
	DT	0	0
	RF	0	0

همسایگی ($k = 5$)، بیز ساده، ماشین بردار پشتیبان (کرل خطی)، درخت تصمیم و جنگل تصادفی ($ntree = 500$) استفاده شده است. جدول ۴ عملکرد مدل‌ها را بر اساس معیارهای دقت (Accuracy) به همراه فاصله اطمینان ۹۵ درصد، کاپای کوهن (Cohen's Kappa) و F1-Macro نشان می‌دهد.

نتایج نشان می‌دهد که الگوریتم‌های ساده‌تر مانند درخت تصمیم (DT) و بیز ساده (NB) عملکرد ضعیف‌تری در تشخیص شدت سوانح داشتند. این موضوع به‌ویژه ناشی از سادگی این مدل‌ها در استفاده از داده‌های متنی است و باعث شده تا این الگوریتم‌ها توانایی کافی برای شناسایی شدت سوانح نداشته باشند و به دلیل ناپایداری در پیش‌بینی کلاس‌های مختلف، دقت چندان بالایی ندارد. الگوریتم KNN عملکرد بهتری نسبت به مدل‌های SVM و Naive Bayes نشان داده است. در مقابل، جنگل تصادفی (Random Forest) بهترین عملکرد را در میان مدل‌های دیگر دارد. این الگوریتم با استفاده از ترکیب چندین درخت تصمیم و بهره‌گیری از رأی‌گیری گروهی توانسته است تا نتایجی پایدارتر ارائه کند. دقت کلی این مدل حدود ۰/۶۹ درصد و همچنین بر اساس معیار کاپا (۰/۴۴۸) و مقدار F1 ماکرو (۰/۴۶۷)، نسبت به سایر مدل‌ها عملکرد بهتری دارد.

نتایج جدول ۵ نشان می‌دهد که عملکرد مدل‌های یادگیری ماشین در طبقه‌بندی شدت سوانح هوایی با توجه به کلاس‌ها بسیار متفاوت است. در میان مدل‌ها جنگل تصادفی (RF)، بر اساس معیارهای درستی (Precision) و F1 عملکرد بهتر یا معادل را برای تمام کلاس‌ها داشته است که نشان‌دهنده توانایی مناسب این مدل در شناسایی نمونه‌های کلاس‌های مختلف است. مدل‌های دیگر مانند DT و NB برای اکثر کلاس‌ها، در قیاس با سایر مدل‌ها عملکرد ضعیف‌تری در طبقه‌بندی داشته‌اند.

همچنین در کلاس‌های نادر مانند «Airprox-Serious Incident» تمام مدل‌ها ناتوان بوده و معیارهای درستی و F1 برابر صفر شده است. به‌طور کلی، این نتایج حاکی از چالش مدل‌ها در شناسایی کلاس‌های کم‌نمونه و عدم تعادل داده‌ها است. این نتایج نشان می‌دهد که الگوریتم‌های پیچیده‌تر و با توانایی یادگیری از نمونه‌های اندک، مانند RF، عملکرد بهتری نسبت به مدل‌های ساده‌تر دارند.

ماتریس درهم‌ریختگی مدل RF در جدول ۶ نشان می‌دهد که این مدل در دسته‌بندی کلاس پر تکرار مانند Accident موفق‌تر بوده، اما در شناسایی کلاس‌های کم‌نمونه‌تر مانند Airprox-Serious Incident همچنان دچار خطا می‌شود. این امر نشان می‌دهد که اگرچه روش‌های کلاسیک یادگیری ماشین می‌توانند تصویری اولیه از قابلیت دسته‌بندی شدت سوانح ارائه دهند، اما برای رسیدن به دقت بالاتر، باید از داده‌های گسترده‌تر و تکنیک‌های پیشرفته‌تر یا مدل‌های پیچیده‌تر استفاده کرد.

ساده‌تر مانند DT و NB توانایی محدودی در تفکیک طبقات داشتند، اما الگوریتم جنگل تصادفی (Random Forest) با دقت ۰/۶۹ درصد و بالاترین مقادیر کاپا و F1، بهترین عملکرد را ارائه داد. این نتیجه نشان می‌دهد که استفاده از مدل‌های ترکیبی (Ensemble Models) و مبتنی بر رأی‌گیری جمعی می‌تواند اثربخش‌تر باشد. با این حال، همچنان چالش‌هایی در شناسایی طبقات کم‌نمونه (مانند Airprox-Serious Incident) باقی مانده است.

از منظر کاربردی، نتایج این پژوهش می‌تواند به سازمان‌های هوانوردی در شناسایی الگوهای پرتکرار، تحلیل عوامل خطرزا و پشتیبانی از تصمیم‌گیری‌های ایمنی کمک کند. به‌ویژه آنکه ترکیب روش‌های آماری و یادگیری ماشین با متن کاوی می‌تواند مسیر تازه‌ای برای تحلیل داده‌های کیفی در صنعت هوایی کشور فراهم آورد.

محدودیت‌های داده‌ای و موضوعی موجود در این پژوهش تأثیر مستقیم و قابل‌توجهی بر نتایج مدل‌سازی موضوعی و طبقه‌بندی شدت سوانح داشته‌اند. نخست، حجم نسبتاً محدود داده‌ها (۱۳۹ گزارش) موجب شد مدل STM نتواند تنوع کامل مضامین موجود در گزارش‌های سوانح را پوشش دهد؛ به‌ویژه موضوعاتی که در تعداد اندکی از گزارش‌ها تکرار شده‌اند، بازنمایی ضعیف‌تری پیدا کردند و در برخی موارد درون موضوعات پرتکرار ادغام شدند. این مسئله باعث کاهش تفکیک‌پذیری موضوعات و از دست رفتن جزئیات معنایی در برخی حوزه‌ها می‌شود. دوم، نامتوازن بودن قابل‌توجه طبقات شدت سوانح باعث شد مدل‌های یادگیری ماشین به سمت طبقات پرتکرار سوگیری پیدا کنند و در تشخیص طبقات کم‌نمونه مانند Serious Incident یا Airprox-Serious Incident عملکرد ضعیف‌تری داشته باشند. گزارش‌هایی که واژگان مشترک میان چند طبقه دارند، باعث خطا در طبقه‌بندی می‌شوند. در مجموع، این محدودیت‌ها سبب شدند دقت مدل‌های طبقه‌بندی کاهش یابد و موضوعات استخراج‌شده از تنوع کمتری برخوردار باشند.

همچنین پژوهشگران آتی می‌توانند با استفاده از مدل‌های پیشرفته‌تر مانند شبکه‌های عصبی به افزایش دقت مدل‌های یادگیری ماشین کمک کنند. همچنین پیشنهاد می‌شود تا سیاست‌گذاران در راستای ذخیره‌سازی و بهره‌برداری از داده‌های سوانح هوایی در پژوهش‌های علمی و دانشگاهی بستر مناسبی را فراهم کنند.

تعارض منافع

هیچ گونه تعارض منافع توسط نویسندگان بیان نشده است.

مراجع

- [1] A. Nanyonga, U. Turhan, and G. Wild, "A systematic review of machine learning analytic methods for

جدول ۶- ماتریس درهم‌ریختگی برای مدل جنگل تصادفی (۵۰۰ درخت) در پیش‌بینی شدت سوانح هوایی.

Table 6. Confusion matrix for Random Forest model (500 trees) in predicting the severity of air accidents.

Reference \ Prediction	Accident	Incident	Serious Incident	Airprox-Serious Incident
Accident	28	9	2	2
Incident	0	5	4	0
Serious Incident	0	0	5	0
Airprox-Serious Incident	0	0	0	0

به‌طور کلی، نتایج بخش دسته‌بندی نشان می‌دهد که محتوای متنی گزارش‌های سوانح می‌تواند به‌عنوان منبعی ارزشمند برای پیش‌بینی نوع و شدت حادثه مورد استفاده قرار گیرد. این موضوع ظرفیت بالای روش‌های متن کاوی و یادگیری ماشین در پشتیبانی از تصمیم‌گیری‌های ایمنی هوایی را تأیید می‌کند.

۵. نتیجه‌گیری

این پژوهش با هدف تحلیل گزارش‌های نهایی سوانح هوایی ایران در بازه‌ی زمانی ۲۰۰۷ تا ۲۰۲۵ با استفاده از روش‌های متن کاوی و الگوریتم‌های یادگیری ماشین بررسی شده است. داده‌ها شامل ۱۳۹ گزارش منتشر شده توسط سازمان هواپیمایی کشوری است که پس از پیش‌پردازش متنی مورد تحلیل و بررسی قرار گرفته‌اند. یافته‌ها نشان دادند که تحلیل متون گزارش‌های سوانح می‌تواند بینش‌های ارزشمندی در خصوص سوانح هوایی ارائه کند.

در بخش نخست، تحلیل فراوانی واژه‌ها مشخص ساخت که بیشترین تمرکز گزارش‌ها بر واژگانی همچون «پرواز»، «خلبان»، «فرودگاه» و «موتور» قرار دارد. این امر نشان‌دهنده اهمیت سه حوزه اصلی عملکرد انسانی، زیرساخت‌های فرودگاهی و مسائل فنی مرتبط با هواپیما در سوانح هوایی است. در ادامه، مدل‌سازی موضوعی با استفاده از الگوریتم STM هشت موضوع کلیدی را آشکار ساخت که مضامینی چون عملکرد موتور، فرآیند تقرب و فرود، آموزش هوانوردی و سوانح بالگردی را شامل می‌شدند. این نتایج بیانگر آن است که سوانح هوایی متأثر از موضوعات مختلفی مانند عوامل فنی و مکانیکی، عملکرد خلبانان و شرایط محیطی هستند.

در بخش دسته‌بندی شدت سوانح، الگوریتم‌های مختلف یادگیری ماشین مورد بررسی قرار گرفتند. نتایج نشان داد که اگرچه الگوریتم‌های

- Journal*, vol. 9, 2023, Art. no. 100358, <https://doi.org/10.1016/j.dajour.2023.100358>.
- [10] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, vol. 116, pp. 48-63, 2019, <https://doi.org/10.1016/j.dss.2018.10.009>.
- [11] R. A. Burnett and D. Si, "Prediction of injuries and fatalities in aviation accidents through machine learning," in *International Conference on Compute and Data Analysis*, Lakeland FL, USA. 2017, pp. 60-68, <https://doi.org/10.1145/3093241.3093288>.
- [12] Y. Xu, Z. Gan, R. Guo, X. Wang, K. Shi, and P. Ma, "Hazard analysis for massive civil aviation safety oversight reports using text classification and topic modeling," *Aerospace*, vol. 11, no. 10, 2024, Art. no. 837, <https://doi.org/10.3390/aerospace11100837>.
- [13] H. Zhang and Q. Wang, "Risk identification model of aviation system based on text mining and risk propagation," *Maintenance and Reliability (Eksploracja i Niezawodność zs)*, vol. 27, no. 1, 2025, <https://doi.org/10.17531/ein/192767>.
- [14] E. Fazeli Sanghani and M. kheirabadi, "Analysis of air accidents in aviation industry," in *3rd International Conference on Science and Engineering*, Istanbul, Turkey, 2016, (in Persian), <https://civilica.com/doc/491596>
- [15] J. Raee, H. R. Zarghami, and A. A. Zand, "Analysis of I.R.I.A.A aviation accidents from 1989 to 2019 using data mining tools," *Military Science and Tactics*, vol. 17, no. 56, pp. 29-50, 2021, (in Persian), <https://doi.org/10.22034/qjmst.2021.136988.1446>.
- [16] "Aircraft Accident Investigation Group (AIG) System," Civil Aviation Organization of Iran, (in Persian), [Online]. Available: <https://aig.cao.ir>
- [17] Annex 13 to the Convention on International Civil Aviation: Aircraft Accident and Incident Investigation, thirteenth ed. ICAO. 2024, [Online]. Available: <https://store.icao.int/en/annex-13-aircraft-accident-and-incident-investigation>
- aviation accident research," *Sci*, vol. 7, no. 3, 2025, Art. no. 124, <https://doi.org/10.3390/sci7030124>.
- [2] Q. Wang, R. Xia, J. Yu, Q. Liu, S. Tong, and Z. Xu, "From text to safety: A novel framework for mining unsafe aviation events using advanced neural network and feature network," *Aerospace*, vol. 11, no. 10, 2024, Art. no. 843, <https://doi.org/10.3390/aerospace11100843>.
- [3] Y. Xing *et al.*, "Discovering latent themes in aviation safety reports using text mining and network analytics," *International Journal of Transportation Science and Technology*, vol. 16, pp. 292-316, 2024, <https://doi.org/10.1016/j.ijtst.2024.02.009>.
- [4] A. Nanyonga, H. Wasswa, U. Turhan, O. Molloy, and G. Wild, "Sequential classification of aviation safety occurrences with natural language processing," in *AIAA Aviation Forum*, San Diego, CA and Online, 2023, Paper 4325, <https://doi.org/10.2514/6.2023-4325>.
- [5] B. Xu and S. Kumar, "A text mining classification framework and its experiments using aviation datasets," in *Workshop of Information Technology and Systems*, Dallas, 2015.
- [6] X. Jing, A. Chennakesavan, C. Chandra, M. V. Bendarkar, M. Kirby, and D. N. Mavris, "BERT for aviation text classification," in *AIAA Aviation Forum*, San Diego, CA and Online, 2023, Paper 3438, <https://doi.org/10.2514/6.2023-3438>.
- [7] Y. Jiao, J. Dong, J. Han, and H. Sun, "Classification and causes identification of Chinese civil aviation incident reports," *Applied Sciences*, vol. 12, no. 21, 2022, Art. no. 10765, <https://doi.org/10.3390/app122110765>.
- [8] V. de Vries, "Classification of aviation safety reports using machine learning," in *International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT)*, Singapore, 2020, <https://doi.org/10.1109/AIDA-AT48540.2020.9049187>.
- [9] D. Ziakkas and K. Pechlivanis, "Artificial intelligence applications in aviation accident classification: A preliminary exploratory study," *Decision Analytics*