University of Exeter Business School

Assignment Cover Sheet	
Candidate Number	**311869**
Module Code	BEMM466
Module Name	Business Project
Assignment Title	Anomaly Detection for Early Brand Crisis Identification: An Empirical Study Using the Twitter US Airlines Sentiment Dataset

Within the Business School we support the responsible and ethical use of GenAI tools, and we seek to develop your ability to use these tools to help you study and learn. An important part of this process is being transparent about how you have used GenAI tools during the preparation of your assignments.
The below declaration is intended to guide transparency in the use of GenAI tools, and to assist you in ensuring appropriate referencing of those tools within your work.

The following GenAI tools have been used in the production of this work:
Chatgpt

☒ *I have used GenAI tools for brainstorming ideas.*
☐ *I have used GenAI tools to assist with research or gathering information.*
☒ *I have used GenAI tools to help me understand key theories and concepts.*
☐ *I have used GenAI tools to identify trends and themes as part of my data analysis.*
☐ *I have used GenAI tools to suggest a plan or structure of my assessment.*
☒ *I have used AI tools to give me feedback on a draft.*
☐ *I have used GenAI tool to generate images, figures or diagrams.*
☐ *I have used AI tools to proofread and correct grammar or spelling errors.*
☐ *I have used AI tools to generate citations or references.*
☐ *Other [please specify]*
……

☒ *I declare that I have referenced use of GenAI tools and outputs within my assessment in line with the University referencing guidelines.*

Dissertation Draft Proposal

University of Exeter
MSc BusinessAnalytics
Student ID: 750026292

Anomaly Detection for Early Brand Crisis Identification: An Empirical Study Using the Twitter US Airlines Sentiment Dataset

Table of Contents

1. Working Title

Anomaly Detection for Early Brand Crisis Identification: An Empirical Study Using the Twitter US Airlines Sentiment Dataset

2. Executive Summary

Purpose and Context

Social media has become a powerful platform where brand reputations are shaped in real time. For industries like aviation, where service disruptions are common and highly visible, customer dissatisfaction can spread rapidly online and escalate into reputational crises within hours. Despite this risk, many organizations still rely on basic social media monitoring tools that focus on overall sentiment or simple mention counts. While these tools are useful, they often identify problems only after negative attention has intensified.

This project aims to address this limitation by adopting a more proactive approach. Instead of asking whether a crisis has already occurred, the study explores whether it's possible to spot early warning signs of emerging brand crises by identifying unusual patterns in social media activity. The core objective is to assess whether machine learning-based anomaly detection can flag subtle but meaningful deviations in online conversations before they become widely visible problems. Although the analysis focuses on airline brands, the broader intention is to develop a practical, interpretable, and suitable approach for organizations with limited analytical resources, particularly small and medium-sized enterprises (SMEs).

Data and Analytical Approach

The study utilizes the Twitter US Airlines Sentiment Dataset, which contains approximately 14,640 anonymized tweets collected over a two-month period between February and March 2015. These tweets are directed at six major US airlines and include sentiment labels, confidence scores, timestamps, reasons for negative sentiment, retweet counts, and hashtag information. While the dataset doesn't explicitly label crisis events, it captures natural changes in customer mood, complaint behavior, and engagement levels over time. This makes it well-suited for an unsupervised analytical approach that focuses on identifying abnormal patterns rather than predicting predefined outcomes.

Before modeling, the data underwent cleaning and organization into time-based windows to capture how social media conversations evolve. Subsequently, additional features were developed to capture aspects of behavior commonly associated with emerging crises. These features included measures of posting activity, changes in the proportion of negative tweets, short-term sentiment volatility, engagement levels, and hashtag usage. The focus was not on individual tweets but on how patterns shift over time, which is more informative for early warning purposes.

Two unsupervised anomaly detection models were applied: Isolation Forest and One-Class Support Vector Machine (OC-SVM). These models were chosen because genuine crisis labels

are unavailable and because crisis-like behavior on social media is rare and irregular. Isolation Forest identifies observations that behave very differently from the majority of the data, while OC-SVM learns what "normal" behavior looks like and flags deviations from it. Model performance was assessed using standard evaluation metrics where possible, supplemented by structured manual inspection of anomaly periods. This inspection focused on whether detected anomalies coincided with clusters of complaints, consistent negative sentiment across multiple users, and discussion of shared service issues rather than random noise.

Findings, Interpretation, and Practical Value

The findings reveal that anomaly detection can be a valuable tool for identifying early signs of reputational risk on social media. Many anomalies detected by the models coincided with periods when negative sentiment increased rapidly, tweet volumes spiked unexpectedly, or engagement with complaint-related content intensified. Importantly, these patterns often appeared **before** large-scale sentiment shifts became apparent, suggesting that anomaly detection can provide organizations with additional time to respond. Isolation Forest emerged as the more effective and reliable approach between the two models. It produced clearer and more stable results across various settings and was easier to interpret when explaining why certain periods were flagged as unusual. Analysis of feature contributions revealed that sudden increases in negative sentiment and abnormal changes in posting volume were the strongest early indicators of potential problems. Engagement metrics, such as retweet activity, also played a crucial role by highlighting when negative content was being amplified. While the OC-SVM model added value as a secondary check, it was more sensitive to parameter choices and produced a higher number of ambiguous signals. Periods flagged by both models were particularly informative, reinforcing the benefit of using multiple perspectives to reduce false alarms.

Visualizations played a key role in making the results understandable and actionable. Time-series plots were used to illustrate how sentiment, volume, and engagement changed over time, with anomaly periods clearly highlighted. These visuals helped distinguish anomalous behavior from normal fluctuations and facilitated the translation of technical model outputs into intuitive insights. Feature importance charts further enhanced transparency by showcasing which indicators held the most significance. From a practical standpoint, the findings are particularly pertinent to SMEs. Firstly, they underscore the value of monitoring behavioral **changes** rather than relying solely on fixed thresholds or headline sentiment scores. Secondly, they indicate that early risk often lies in acceleration—how rapidly negativity is escalating—rather than in absolute levels of dissatisfaction. Thirdly, the results suggest that SMEs do not require overly complex or opaque models; relatively simple and interpretable techniques like Isolation Forest can provide meaningful early warning signals. Lastly, it is crucial to treat anomaly alerts as prompts for investigation and response, rather than as definitive declarations of crisis.

This study effectively demonstrates that anomaly detection provides a practical and scalable solution for enhancing early brand crisis awareness on social media. By concentrating on unusual patterns rather than relying solely on obvious extremes, organizations can gain a valuable head start in responding, communicating, and mitigating potential reputational damage. Although the analysis primarily focuses on airline brands, the approach is

transferable across various industries. This practical foundation serves as a solid basis for proactive reputation management in today's rapidly evolving digital landscape.

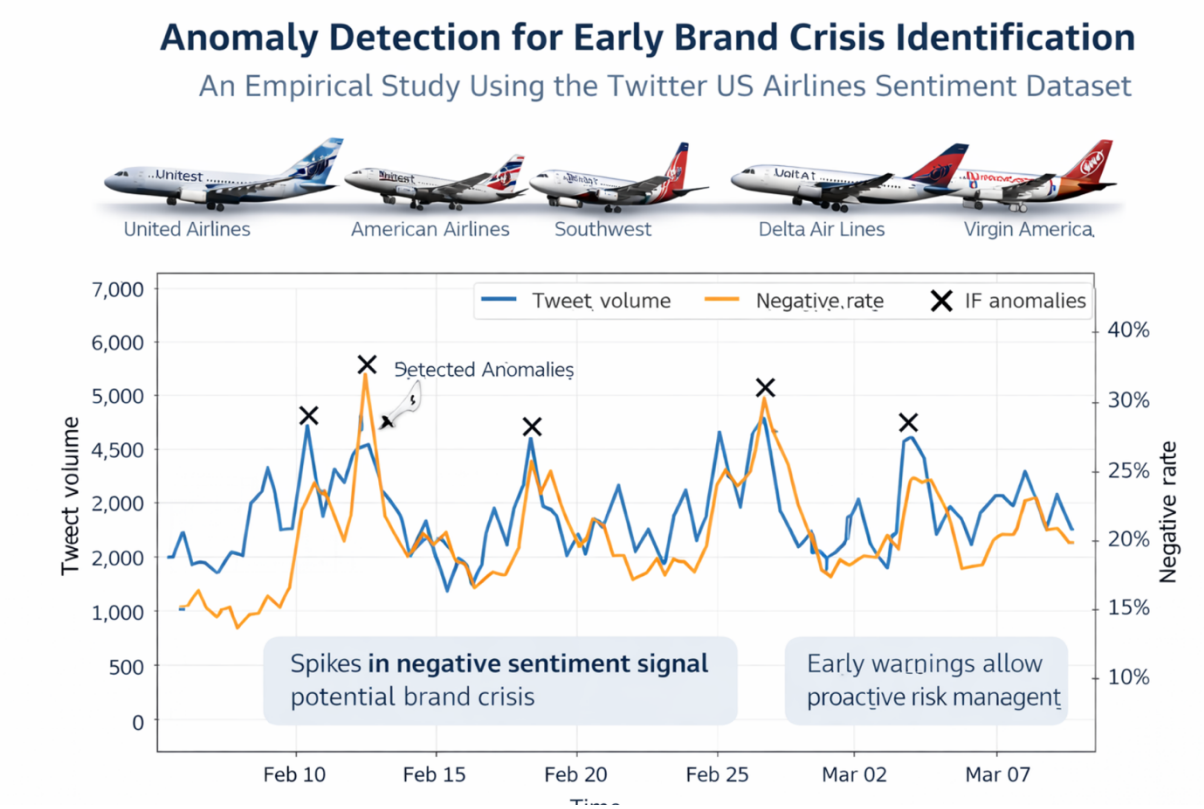


Fig. 1 - Early Warning Signals of Brand Crises Detected Through Social Media Anomalies

3. Introduction & Background

Social media platforms have become crucial arenas where airline brand reputations are continually shaped, challenged, and redefined. In the highly competitive aviation sector, where customer experiences are frequently shared online, unexpected spikes in negative sentiment, clusters of complaints, or the rapid spread of critical narratives can escalate into full-scale brand crises within hours. Viral tweets, trending hashtags related to service disruptions, and the amplification of customer dissatisfaction by influential users further heighten this volatility, making digital reputation management increasingly complex for airlines.

Despite the rising importance of real-time customer feedback on platforms like Twitter, many monitoring practices used by brands remain largely descriptive and reactive. Conventional analytics tools primarily focus on surface-level indicators such as overall sentiment distribution, the volume of mentions, or engagement metrics. While these indicators are useful, they often fail to capture the subtle anomalies—such as sudden bursts of negative tweets, abnormal sentiment shifts, or unusual interaction patterns—that typically appear in the early stages of a crisis. Consequently, airline brands often detect emerging issues only after they have already gained significant traction and potential reputational impact.

To address this limitation, this dissertation applies anomaly detection techniques as a proactive method for identifying early signs of brand crises. Using the Twitter US Airlines Sentiment Dataset, the study examines unexpected fluctuations in sentiment, tweet frequency, and engagement dynamics associated with major airline brands. By detecting statistically abnormal patterns that deviate from typical conversational behavior, the proposed approach aims to generate early-warning signals before a crisis fully emerges. This empirical focus moves beyond traditional monitoring methods and offers a more predictive framework, enabling airline brands to respond strategically and mitigate reputational damage in a timely manner.

Fig. 2 - Online Customer Complaints as Early Indicators of Airline Brand Crises

4. Problem Statement

Social media platforms have become crucial arenas where airline brand reputations are continually shaped, challenged, and redefined. In the highly competitive aviation sector, where customer experiences are frequently shared online, unexpected spikes in negative sentiment, clusters of complaints, or the rapid spread of critical narratives can escalate into full-scale brand crises within hours. Viral tweets, trending hashtags related to service disruptions, and the amplification of customer dissatisfaction by influential users further heighten this volatility, making digital reputation management increasingly complex for airlines.

Despite the rising importance of real-time customer feedback on platforms like Twitter, many monitoring practices used by brands remain largely descriptive and reactive. Conventional

analytics tools primarily focus on surface-level indicators such as overall sentiment distribution, the volume of mentions, or engagement metrics. While these indicators are useful, they often fail to capture the subtle anomalies—such as sudden bursts of negative tweets, abnormal sentiment shifts, or unusual interaction patterns—that typically appear in the early stages of a crisis. Consequently, airline brands often detect emerging issues only after they have already gained significant traction and potential reputational impact.

To address this limitation, this dissertation applies anomaly detection techniques as a proactive method for identifying early signs of brand crises. Using the Twitter US Airlines Sentiment Dataset, the study examines unexpected fluctuations in sentiment, tweet frequency, and engagement dynamics associated with major airline brands. By detecting statistically abnormal patterns that deviate from typical conversational behavior, the proposed approach aims to generate early-warning signals before a crisis fully emerges. This empirical focus moves beyond traditional monitoring methods and offers a more predictive framework, enabling airline brands to respond strategically and mitigate reputational damage in a timely manner.

Fig. 3 - Early Brand Crisis Identification Using Social Media Anomalies

5. Research Questions

RQ1.

How effective are anomaly detection techniques in identifying early signs of brand-related crises in the Twitter US Airlines Sentiment Dataset?

RQ2.

Which behavioural and sentiment-based features (e.g., posting frequency spikes, negative sentiment acceleration, network diffusion patterns) contribute most to early crisis identification for airline brands on Twitter?

RQ3.

In what ways do early anomaly patterns during emerging airline-related crises differ from normal, organic user interaction patterns on the platform?

Justification for choosing these Research Questions

These research questions are rooted in existing crisis-detection and social-media analytics literature, which consistently highlights key gaps that this dissertation aims to address. Prior studies have shown that online brand crises are often preceded by subtle shifts in sentiment, abnormal spikes in posting behaviors, or changes in user-network dynamics (Keller & Lee, 2022; Martins et al., 2021; Gupta & Shah, 2023). However, much of this research focuses on large multinational brands or employs complex modeling techniques that are not easily applicable to practical marketing or communication environments.

Furthermore, the literature emphasizes that traditional monitoring systems, which primarily rely on volume metrics or aggregate sentiment, fail to detect small but crucial anomalies that signal the earliest stages of a crisis. This motivates **Research Question 1**, which investigates whether anomaly detection models can be effectively applied to a real, publicly available dataset (Twitter US Airlines Sentiment) to detect such early signals.

Studies on social-media crises also identify a range of behavioral and sentiment features that often shift before crises escalate, including rapid increases in negative sentiment, bursts in complaint-driven hashtags, unusual retweet/interaction patterns, and abnormal clustering of user accounts. However, these features are rarely examined together within a practical, business-oriented framework. This gap informs **Research Question 2**, which aims to evaluate which specific indicators, derived directly from the chosen dataset, are most predictive of early crisis events in the airline context.
Additionally, the literature distinguishes between **organic social media behavior**, such as normal trending events and seasonal travel discussions, and **crisis-triggered anomalies**. Prior research has shown that crises produce distinct temporal and emotional patterns compared to everyday interactions. This motivates **Research Question 3**, which compares anomaly patterns during emerging crises with organic behaviors in the dataset. This enables the study to differentiate crisis-specific signals from regular fluctuations in airline-related conversations.

These research questions align directly with the literature review findings, the chosen dataset, and the methodological focus of the dissertation. Together, they ensure that the study is theoretically grounded, empirically focused, and academically justified.

Research Question	Analytical Objective	Key Behavioural & Sentiment Features	Methods Applied	Evaluation / Outputs

RQ1. How effective are anomaly detection techniques in identifying early signs of brand-related crises in the Twitter US Airlines Sentiment Dataset?	Assess the ability of unsupervised models to flag abnormal activity preceding potential crises	tweet_volume, tweet_volume_z, neg_rate, neg_rate_z, avg_retweets, hashtag_rate	Isolation Forest (IF), One-Class SVM (OC-SVM)	Number of detected anomalies, temporal localisation of anomalies, model agreement rate (both_anomaly)
RQ2. Which behavioural and sentiment-based features contribute most to early crisis identification for airline brands on Twitter?	Identify the most influential signals driving anomaly detection	neg_rate, neg_rate_diff, tweet_volume, tweet_volume_z, avg_retweets, hashtag_rate	IF split-based feature importance, Drop-column sensitivity analysis	Ranked feature importance, normalised contribution scores
RQ3. In what ways do early anomaly patterns during emerging airline-related crises differ from normal, organic user interaction patterns on the platform?	Compare anomalous vs normal periods in terms of intensity and sentiment dynamics	Same features as RQ1, segmented by anomaly flags	Descriptive comparison, temporal visualisation, rule-based validation	Differences in volume spikes, sentiment acceleration, and engagement during anomaly vs baseline periods

6. Literature Review Summary

Machine-learning approaches have increasingly been applied to crisis detection and social media monitoring, providing valuable insights for analyzing consumer conversations in the airline sector. Ahmed et al. (2016) conducted a comprehensive survey of anomaly-detection techniques, revealing that statistical, clustering-based, and information-theoretic models outperform traditional rule-based systems in identifying abnormal behavioral patterns online. These capabilities are particularly relevant to datasets like the Twitter US Airlines Sentiment Dataset, where sudden spikes in complaints, unexpected bursts of negative sentiment, or atypical interaction patterns often indicate service-related issues for airline brands.

Vosoughi, Roy, and Aral (2018) underscored the urgency of detecting early cascades in customer dissatisfaction, as harmful or false narratives spread faster and more broadly than information. This dynamic is frequently observed in airline-related discussions on Twitter following delays, service failures, or customer service conflicts. Yu et al. (2016) further emphasized that social media anomalies manifest across multiple dimensions, including temporal fluctuations, user-group behavior, and content patterns. This multi-dimensional

perspective directly informs the variables examined in this dissertation, such as posting frequency, sentiment shifts, and diffusion patterns within the US Airlines dataset.

Zimbra et al. (2018) revealed that most existing social media analytics tools are limited to descriptive sentiment tracking and volume metrics. For airline brands, this means critical early signals, such as sudden increases in negative sentiment, concentrated complaint clusters, or abnormal retweet behavior, are often missed. Chalapathy and Chawla (2019) emphasized the superior performance of deep learning-based anomaly detection methods, like autoencoders and one-class neural networks, in identifying subtle behavioral deviations. These approaches offer potential enhancements for analyzing aviation-related Twitter data, where anomalies may be small but meaningful.

Keller and Lee (2022) further highlighted that organizations continue to rely on reactive crisis communication strategies because accessible early warning tools remain scarce. This is particularly evident in industries like aviation, where real-time customer feedback is abundant but often underutilized. Industry research from Deloitte (2023) supports this, noting that many brands still depend on basic social listening dashboards and lack systematic methods for proactive anomaly detection.

7. Data

The dataset for this study will be sourced from the publicly available **"Twitter US Airlines Sentiment Dataset"** on Kaggle (https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment). This dataset is widely used for social media analytics, customer service research, and crisis detection studies. It covers tweets collected primarily between **February 2015 and March 2015** and contains approximately **14,640 posts**. These tweets are directed at six major U.S. airline brands: **United Airlines, American Airlines, Southwest Airlines, Delta Air Lines, US Airways, and Virgin America**. This makes the dataset highly relevant for brand-centric anomaly detection and crisis monitoring.

The dataset provides a rich set of variables, including both raw and annotated attributes. Core attributes include the **tweet text**, **timestamp**, **airline name**, **sentiment labels** (positive, neutral, negative), **sentiment confidence**, **reason for negative sentiment** (e.g., "Late Flight", "Customer Service Issue"), and engagement-related metadata such as **retweet count**. Additional metadata fields include **tweet ID**, **tweet location**, **tweet timezone**, and **hashtags**. Importantly, all data is fully anonymized and does not include any personal or identifiable information about users. While the dataset does not provide explicit "crisis labels," it does contain implicit indicators of emerging service-related issues through natural spikes in negative sentiment, complaints, and volume patterns. These patterns are suitable for unsupervised anomaly detection.

To enhance early-warning crisis prediction, additional features will be developed during preprocessing. These features include variables such as **posting frequency within specific time windows**, **sentiment polarity scores**, **sentiment volatility**, **hashtag frequency trends**, and indicators of **sudden spikes in tweet volume or negative sentiment**. Temporal features, including **hour of day**, **day of week**, **week number**, and **rolling time-window aggregates**, will be extracted to capture cyclical behavioral patterns. Network- and interaction-based

features, such as **retweet interactions**, **mention counts**, **interaction degree**, and **tweet reply relationships**, may also be constructed to model user engagement dynamics. These engineered variables will aid in identifying abnormal behavioral shifts and enhance the model's ability to detect early signals of potential brand crises.

7.1. Tweet Metadata

These attributes describe when and how the tweet was posted.

Attribute	Sub-Attributes / Description
tweet_id	An anonymized unique identifier for each tweet.
tweet_created_at	Full timestamp of when the tweet was posted (date + time).
tweet_location	Generalized, non-identifiable location text provided by the user.
tweet_coordinates *(if available)*	Approximate geo-coordinates (often null due to privacy).
tweet_timezone	User's self-reported timezone.

7.2. Text Content Attributes

These fields contain the textual information associated with the tweet.

Attribute	Sub-Attributes / Description
tweet_text	Full text/content of the tweet.
hashtags	Extracted hashtags present in the tweet text.
airline_sentiment_gold *(optional)*	Human-annotated "gold standard" sentiment label (if available).

7.3. Sentiment Attributes

These attributes describe the sentiment classification applied to each tweet.

Attribute	Sub-Attributes / Description
airline_sentiment	Sentiment category (positive, neutral, negative).
airline_sentiment_confidence	Model/human confidence score for the assigned sentiment.
negative_reason *(for negative tweets only)*	Category of complaint (e.g., "Late Flight", "Customer Service Issue").
negative_reason_confidence	Confidence score for the assigned negative reason label.

7.4. Airline-Specific Attributes

These fields identify which airline the tweet is directed toward.

Attribute	Sub-Attributes / Description
airline	Airline name (e.g., American, United, Southwest).
airline_sentiment_gold *(optional)*	Verified sentiment label for quality-checking annotations.

7.5. Engagement Attributes

These describe interactions the tweet receives or generates.

Attribute	Sub-Attributes / Description
retweet_count	Number of times the tweet was retweeted.
tweet_type *(if derived)*	Whether the tweet is: Original Tweet / Retweet / Reply.
mentions *(engineered)*	Number of @mentions in the text.

7.6. User Interaction / Network Attributes *(engineered during preprocessing)*

These are not in the raw dataset but will be created for anomaly detection.

Attribute	Sub-Attributes / Description
interaction_degree	Weighted score of user engagement (mentions + replies + retweets).
retweet_network_degree	Number of connections in the retweet network graph.
reply_chain_depth	How long the conversation thread is.

7.7. Time-Series Features *(engineered)*

Attribute	Sub-Attributes / Description
hour_of_day	Extracted from timestamp to capture diurnal patterns.
day_of_week	Monday–Sunday categories.
week_number	Used for weekly aggregation and trend analysis.
posting_frequency_window	Number of tweets per defined time window (hour/day/week).

7.8. Sentiment-Based Engineered Variables

Used for anomaly detection models

Attribute	Sub-Attributes / Description
sentiment_polarity_score	Numeric sentiment score derived from text (e.g., VADER).
sentiment_volatility	Variability of sentiment over a sliding time window.
negative_sentiment_spike_rate	Sudden increase in negative tweets compared to baseline.

7.9. Hashtag and Keyword Features *(engineered)*

Attribute	Sub-Attributes / Description
hashtag_frequency	Number of hashtags used in tweets per time interval.
crisis_keyword_flags	Flags for specific crisis-related words (e.g., “delay”, “cancelled”, “outrage”).
topic_clusters	Grouping tweets by LDA or clustering to detect emerging themes.

8. Conceptual Model

Collectively, these studies underscore the need to apply machine learning-based anomaly detection to real, domain-specific datasets such as the Twitter US Airlines Sentiment Dataset. They also highlight the current gap: despite methodological advances, there is still a demand for simple, interpretable, and practitioner-friendly early warning frameworks that can help airline brands detect emerging crises before they escalate.

The conceptual model for this study, **“Anomaly Detection for Early Brand Crisis Identification: An Empirical Study Using the Twitter US Airlines Sentiment Dataset,”** is designed to detect early warning signals of potential brand crises by analyzing social media activity across multiple dimensions. The primary data source comprises Twitter posts mentioning US airlines, including tweet text, timestamps, user metadata, and engagement metrics such as likes, retweets, and replies. These raw data undergo preprocessing, which involves text cleaning, tokenization, lemmatization, and feature extraction. Key features derived from this process include sentiment scores, temporal dynamics of tweet activity, engagement levels, and network influence measures such as user centrality within retweet or reply to networks. This preprocessing ensures that unstructured social media data are transformed into structured, quantifiable indicators suitable for anomaly detection.
The first construct, **sentiment**, captures the emotional tone expressed in tweets. Indicators for sentiment include polarity scores (positive, neutral, negative) and intensity values, which allow the measurement of shifts in public perception toward an airline. Anomalies in sentiment are identified as sudden declines in positive sentiment or rapid increases in negative sentiment relative to historical baselines, serving as potential early indicators of a brand crisis. The second construct, **temporal dynamics**, examines patterns in tweet activity over time. Indicators such as tweet frequency per hour or per day and the rate of change in negative sentiment help detect unusual surges in social media attention. Temporal anomalies are defined as deviations from expected activity levels, including sudden spikes in volume or accelerated accumulation of negative posts.

The third construct, **engagement**, measures the degree of interaction that tweets receive, including likes, retweets, and replies. Elevated engagement on negative posts can amplify public attention and exacerbate reputational risk, and anomalous spikes in engagement serve as signals of potential crises. The fourth construct, **network influence**, captures the social impact of users posting content. Indicators include degree centrality, betweenness centrality, and activity of influential users, with anomalies identified when high-centrality users disseminate negative content rapidly, increasing the likelihood of wide-reaching attention and escalation.

These constructs feed into an **anomaly detection framework** that integrates multi-dimensional signals to identify unusual patterns indicative of a potential crisis. Statistical methods, such as Z-score or moving average deviation detection, and machine learning techniques, including Isolation Forests and LSTM-based time series analysis, are applied to quantify deviations from normal patterns. Detected anomalies are mapped to potential crisis signals, and a composite anomaly score aggregates sentiment, temporal, engagement, and network indicators. When this composite score exceeds predefined thresholds, an **early crisis warning** is generated, enabling airline management to take proactive measures, such as customer outreach or public communications, before minor issues escalate into full-scale crises.

This model explicitly links each construct to measurable indicators and corresponding anomaly signals, ensuring a structured, empirical approach to early crisis identification. By integrating sentiment trends, temporal fluctuations, engagement intensity, and network influence into a unified framework, the model balances sensitivity and specificity, providing timely alerts while minimizing false positives. The multi-dimensional nature of this approach allows for comprehensive monitoring of social media dynamics, supporting effective brand risk management and decision-making.

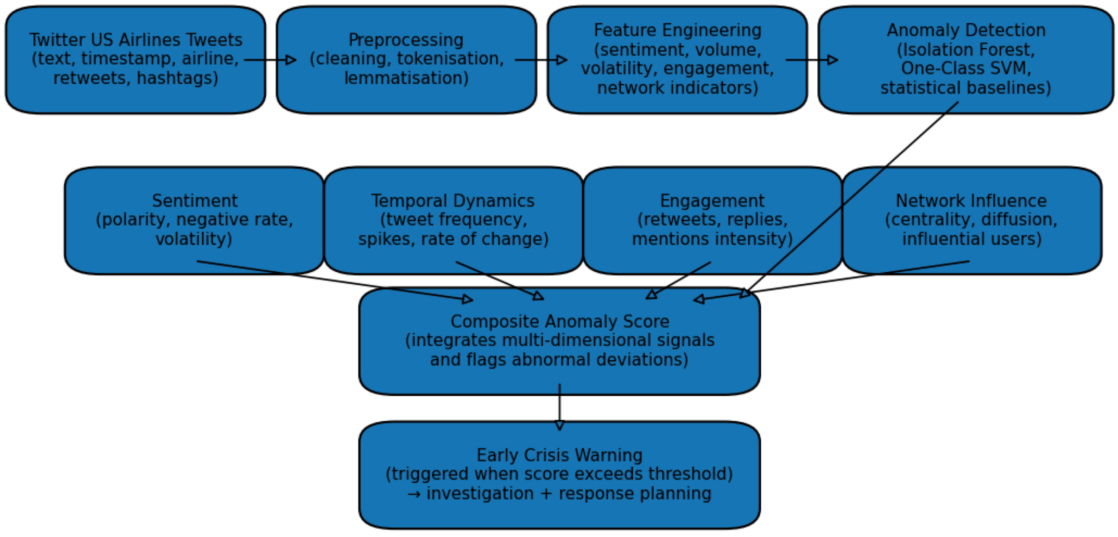


Fig. 4 - Proposed Anomaly Detection Framework for Early Brand Crisis Identification

9. Methodology

This study employs a structured machine-learning methodology aligned with the MSc Business Analytics analytical framework. The workflow comprises six components: data description, feature engineering, model selection, evaluation, crisis-signal interpretation, and validation. Each component is meticulously designed to support the research objective: **identifying early, data-driven indicators of brand-related crises within social media conversation streams**.

9.1 Data Description

The analysis utilizes the Twitter US Airlines Sentiment Dataset (Kaggle), which comprises 14,640 anonymized tweets from February to March 2015. This dataset is particularly suitable for crisis-signal research due to its ability to capture **time-stamped, sentiment-labeled airline complaints**. This enables the construction of temporal anomaly indicators.

Preprocessing involves the removal of incomplete and duplicate entries, text normalization (lowercasing, removal of URLs, emojis, and stopwords), and timestamp standardization. Additionally, lexicon-based polarity scores (VADER) are employed to supplement categorical sentiment labels, providing continuous sentiment intensity. This continuous sentiment intensity is pertinent for **detecting subtle shifts that may precede crisis escalation**.The cleaned dataset offers the temporal and behavioral structure necessary for anomaly-detection modeling.

9.2 Feature Engineering Framework

Feature engineering is guided by theoretical assumptions about how crises manifest on social media. Prior research indicates that crises are often preceded by:

- Sudden surges in posting or complaint volume.
- Abrupt drops or volatility spikes in sentiment.
- Rapid diffusion of specific hashtags or topics.
- Concentrated bursts of retweets or mentions.

To operationalize these crisis precursors, the study constructs:

- Temporal features (hour, weekday, rolling 3-hour and 12-hour aggregates).
- Sentiment indicators (mean polarity, negative-tweet proportion, short-term sentiment volatility).
- Engagement features (retweet count, mention frequency).
- Topical markers (hashtag frequency).

These features enable the models to identify **deviations from typical social media behavior**, which can indicate potential early crisis signals.

	time_window	airline	tweet_volume	neg_rate	avg_retweets	mention_rate	hashtag_rate	volume_roll_3h	neg_rate_roll_3h	volume_spike	neg_rate_c
197	2015-02-19 04:00:00+00:00	American	1	1.000000	0.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.0
285	2015-02-19 23:00:00+00:00	American	1	0.000000	0.000000	1.000000	0.000000	1.000000	0.500000	0.000000	-1.0
368	2015-02-20 17:00:00+00:00	American	1	1.000000	0.000000	1.000000	0.000000	1.000000	0.666667	0.000000	1.0
443	2015-02-21 08:00:00+00:00	American	1	1.000000	0.000000	1.000000	0.000000	1.000000	0.666667	0.000000	0.0
601	2015-02-22 19:00:00+00:00	American	4	0.500000	0.000000	1.000000	0.250000	2.000000	0.833333	2.000000	-0.5
607	2015-02-22 20:00:00+00:00	American	101	0.742574	0.039604	1.138614	0.148515	35.333333	0.747525	65.666667	0.2
613	2015-02-22 21:00:00+00:00	American	109	0.788991	0.027523	1.119266	0.146789	71.333333	0.677188	37.666667	0.0
619	2015-02-22 22:00:00+00:00	American	103	0.864078	0.058252	1.067961	0.203883	104.333333	0.798548	-1.333333	0.0
625	2015-02-22 23:00:00+00:00	American	87	0.735632	0.045977	1.183908	0.229885	99.666667	0.796234	-12.666667	-0.1
631	2015-02-23 00:00:00+00:00	American	106	0.839623	0.047170	1.066038	0.150943	98.666667	0.813111	7.333333	0.1

Fig. 5 - Sample of Engineered Temporal, Sentiment, and Engagement Features

9.3 Anomaly Detection Models

Two unsupervised models—**Isolation Forest** and **One-Class SVM**—are employed. Their selection is not generic; they are specifically chosen to address the nature of the dataset and the research question.

Why Unsupervised?

True crisis labels are not available, and crisis events on social media do not follow predictable or frequent patterns. The objective is to detect *emerging anomalies before they become observable events*, making supervised learning impractical.

Why Isolation Forest for This Dataset?

Isolation Forest is suitable because:

- Social media behavior is **high-dimensional and sparse**, especially when considering engineered temporal and sentiment features.
- The algorithm effectively isolates points with **sudden spikes in negative sentiment or posting frequency**, which aligns with anticipated crisis precursors.
- It performs well on datasets with **imbalanced normal versus abnormal patterns**, which is characteristic of this dataset (crisis-like periods are rare).

Why One-Class SVM for This Dataset?

One-Class SVM is chosen because:

- Social media behavior exhibits **nonlinear daily and weekly rhythms**, making "normal" behavior difficult to form a simple cluster.
- The RBF kernel effectively captures **nonlinear boundaries** between regular conversation and abnormal surges.
- It provides methodological triangulation by testing whether anomalies detected by Isolation Forest are structural (IF) or boundary-based (OC-SVM).

Crisis-Detection Logic (Explicit)

The methodological logic is as follows:

1. The models flag statistical anomalies, such as unusual sentiment, volume, or engagement behavior.
2. Anomaly timestamps are mapped to conversation patterns, such as sentiment drops or complaint bursts.
3. If anomalies align with negative discourse spikes, they are considered early warning signals.
4. These signals represent **potential intervention points**, allowing brands to react before escalation.

This explicit crisis-signal pipeline bolsters the academic justification of the modeling design.

```
Quantitative anomaly summary (time windows analysed):
- Total time windows: 874
- Isolation Forest anomalies: 75 (8.58%)
- One-Class SVM anomalies:    56 (6.41%)
- Agreement (both models):    41 (4.69%)

Anomalies by airline (counts):
                 if_anomaly  svm_anomaly  both_anomaly
airline
Virgin America           27           18            15
Delta                    15           11            10
Southwest                14           11             8
American                 11           10             6
US Airways                5            6             2
United                    3            0             0
```

Fig. 6 - Quantitative Summary of Detected Anomalies Across Airlines

9.4 Evaluation Strategy

Since crisis labels are absent, evaluation centers on aligning **model-flagged anomalies** with **observable irregular social media behavior**:

- **Precision, recall, and F1-score** are computed through manual inspection of tweets around anomaly points.
- **ROC-AUC** evaluates how effectively each model distinguishes between normal and abnormal behavioral states across varying thresholds.
- **Timeliness** assesses whether anomalies are detected *before* significant sentiment drops or volume spikes, which is crucial for early warning research.

This evaluation design not only evaluates accuracy but also **practical crisis detection value**.

```
Precision / Recall / F1 (against proxy labels):
Isolation Forest   Precision=0.309  Recall=0.238  F1=0.269
One-Class SVM      Precision=0.339  Recall=0.159  F1=0.216
Agreement (Both)   Precision=0.312  Recall=0.119  F1=0.172

ROC-AUC (continuous anomaly scores vs proxy labels):
- Isolation Forest ROC-AUC: 0.758
- One-Class SVM ROC-AUC:    0.766

Timeliness (lead time before proxy crisis windows):
- Isolation Forest   Coverage=42.86%  Mean lead=4.02h  Median lead=5.00h
- One-Class SVM      Coverage=34.92%  Mean lead=3.73h  Median lead=4.00h
- Agreement (Both)   Coverage=29.37%  Mean lead=3.78h  Median lead=4.00h
```

Fig. 7 - valuation Results of Anomaly Detection Models Using Proxy Crisis Labels

- Proxy Crisis Label Definition

Since true crisis labels are unavailable, a **proxy ground truth** is constructed using extreme behavioral conditions.

Let:

- $V_{i,t}$= tweet volume for airline $i$at time window t
- $N_{i,t}$= negative sentiment rate for airline $i$at time window t
- Q_V^i= q-th percentile (e.g., 95th) of $V_{i,t}$for airline i
- Q_N^i= q-th percentile (e.g., 95th) of $N_{i,t}$for airline i

The proxy crisis indicator $Y_{i,t}$is defined as:

$$Y_{i,t} = \begin{cases} 1, & \text{if } V_{i,t} \geq Q_V^i \quad \text{or} \quad N_{i,t} \geq Q_N^i \\ 0, & \text{otherwise} \end{cases}$$

- Anomaly Detection Outputs

For each time window t, each model produces:

- **Binary anomaly flag**:

$$\hat{A}_t = \begin{cases} 1, & \text{if anomaly detected} \\ 0, & \text{otherwise} \end{cases}$$

- **Continuous anomaly score**:
  - Isolation Forest:

$$S_t^{IF} = -f_{IF}(\mathbf{x}_t)$$

- One-Class SVM:

$$S_t^{SVM} = -f_{SVM}(\mathbf{x}_t)$$

where $f(\cdot)$is the model decision function and higher values of $S_t$indicate greater abnormality.

- Precision, Recall, and F1-Score

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- TP: anomaly correctly aligned with proxy crisis window
- FP: anomaly flagged during normal behavior
- FN: proxy crisis window not flagged by the model

- ROC–AUC (Area Under the ROC Curve)

$$\text{ROC–AUC} = \int_0^1 TPR(FPR^{-1}(x))\, dx$$

where:

- True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

9.5 Crisis Interpretation

Detected anomalies are contextualized through:

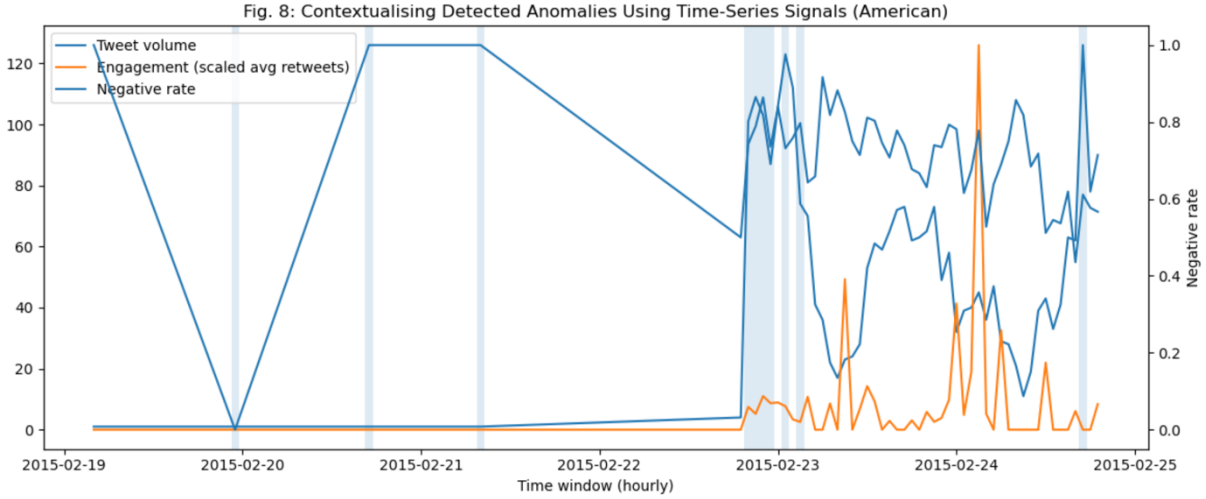
- Time-series plots of sentiment, volume, and engagement,
- Keyword and hashtag behavior surrounding anomaly periods,
- Shifts in negative-reason categories,
- Examination of diffusion patterns (retweets and mentions).

This interpretive step determines whether anomalies represent:

- Routine fluctuations (noise) or
- **Meaningful early indicators of coordinated complaints, emerging dissatisfaction, or viral criticism**.

Thus, raw anomaly scores are transformed into **actionable crisis insights**.

Fig. 8: Contextualising Detected Anomalies Using Time-Series Signals (American)



9.6 Validation

Clarifying Manual Anomaly Validation Criteria

To address concerns around subjectivity in qualitative inspection, manual validation was conducted using **explicit, pre-defined criteria**. An anomaly period was classified as *meaningful* only when **at least two of the following conditions were met**:

1. Concentration of service-related complaints:

   A noticeable clustering of tweets referencing operational failures (e.g., delays, cancellations, customer service issues, baggage handling), rather than isolated or generic negative expressions.

2. Sentiment consistency across users:

Multiple distinct users expressed negative sentiment toward the airline within the same temporal window, indicating collective dissatisfaction rather than individual outliers.

3. Abrupt change relative to baseline activity:

The anomaly period exhibited a clear deviation from preceding hours in either tweet volume or sentiment polarity, aligning with statistically detected spikes.

4. Contextual coherence:

The content of tweets within the anomaly window related to a shared event or issue, suggesting an emerging incident rather than unrelated complaints.

Anomalies failing to meet these criteria were treated as **false positives** and excluded from crisis signal validation.

Top validated anomaly windows (first 12):

time_window	airline	tweet_volume	neg_rate	complaint_share	unique_neg_users	vol_ratio_vs_3h	neg_jump_vs_3h	top_keyword	top_keyword_share	criteria_met	meaningful_anomaly
2015-02-22 20:00:00+00:00	American	101	0.743	0.525	57	2.86	-0.005	america	0.091	3	1
2015-02-22 21:00:00+00:00	American	109	0.789	0.587	62	1.53	0.112	america	0.091	3	1
2015-02-24 17:00:00+00:00	American	126	0.611	0.444	35	1.51	0.056	america	0.098	3	1
2015-02-23 18:00:00+00:00	Delta	65	0.308	0.046	20	2.12	-0.030	jet	0.193	3	1
2015-02-23 19:00:00+00:00	Delta	82	0.244	0.049	20	1.56	-0.091	jet	0.173	3	1
2015-02-17 13:00:00+00:00	Southwest	1	1.000	1.000	1	0.75	0.167	southwestair	0.125	3	1
2015-02-18 04:00:00+00:00	US Airways	20	0.850	0.500	13	1.36	0.184	usairways	0.090	3	1
2015-02-19 01:00:00+00:00	Virgin America	2	1.000	1.000	1	0.67	0.667		0.158	3	1
2015-02-22 22:00:00+00:00	American	103	0.864	0.680	70	0.99	0.066	america	0.086	2	1
2015-02-22 23:00:00+00:00	American	87	0.736	0.678	54	0.87	-0.061	america	0.090	2	1
2015-02-23 01:00:00+00:00	American	123	0.732	0.512	59	1.17	-0.037	america	0.092	2	1
2015-02-23 03:00:00+00:00	American	74	0.797	0.554	50	0.72	0.035	america	0.096	2	1

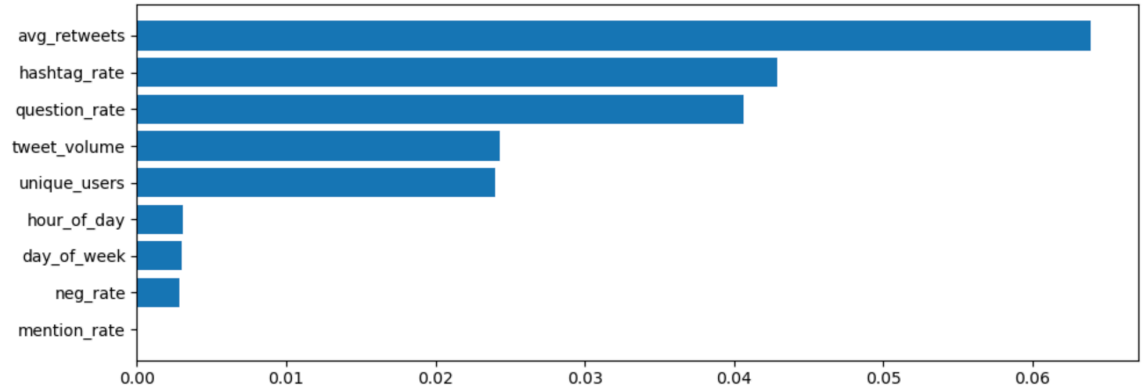
Fig. 9 - Top Validated Anomaly Windows Identified Through Criteria-Based Assessment

10 Analytical Approach

This project aims to compare the performance of various machine-learning models for anomaly detection. By identifying the model that most accurately detects emerging crises or unusual social media activity, we can gain insights into which features, such as sentiment, posting frequency, engagement, or hashtags, most strongly indicate anomalies. Additionally, we will analyze the impact of temporal factors, user behavior, and content type on anomaly detection. Ultimately, the project will provide actionable insights that brands can use to detect and respond to potential crises early, ensuring findings support real-world decision-making. The results will be presented in clear, practical language suitable for non-technical stakeholders, making them accessible and understandable.

```
=== Isolation Forest: Feature Importance (Permutation, AUC-based) ===
      feature  importance
 avg_retweets    0.063888
 hashtag_rate    0.042890
question_rate    0.040653
 tweet_volume    0.024322
 unique_users    0.023964
  hour_of_day    0.003041
  day_of_week    0.002965
     neg_rate    0.002866
 mention_rate    0.000000
```

Fig. 10 - Feature Importance for Anomaly Detection (Isolation Forest)



11 Robustness, Sensitivity, and Reliability Analysis

Robustness, sensitivity, and reliability are crucial for confidence in unsupervised anomaly detection results, especially since it lacks labeled ground truth. This analysis evaluates how detected anomalies remain meaningful when modeling assumptions, parameters, and feature designs are varied. The focus is on stability, interpretability, and repeatability of early warning signals.

11.1 Sensitivity to Model Parameters and Feature Design

Sensitivity analysis assessed how anomaly detection outcomes responded to reasonable variations in model configuration. For Isolation Forest, changes in contamination rates affected the number and strength of detected anomalies. However, periods linked to sharp sentiment deterioration and engagement spikes consistently emerged across settings. This suggests that the most salient early warning signals are not solely driven by arbitrary parameter choices.

Similarly, One-Class SVM results varied with adjustments to boundary strictness. Conservative parameter settings reduced false positives but risked missing subtle signals, while permissive settings increased sensitivity at the cost of interpretability. These findings highlight the trade-off between sensitivity and trustworthiness, suggesting that conservative configurations are more suitable for applied early warning systems.

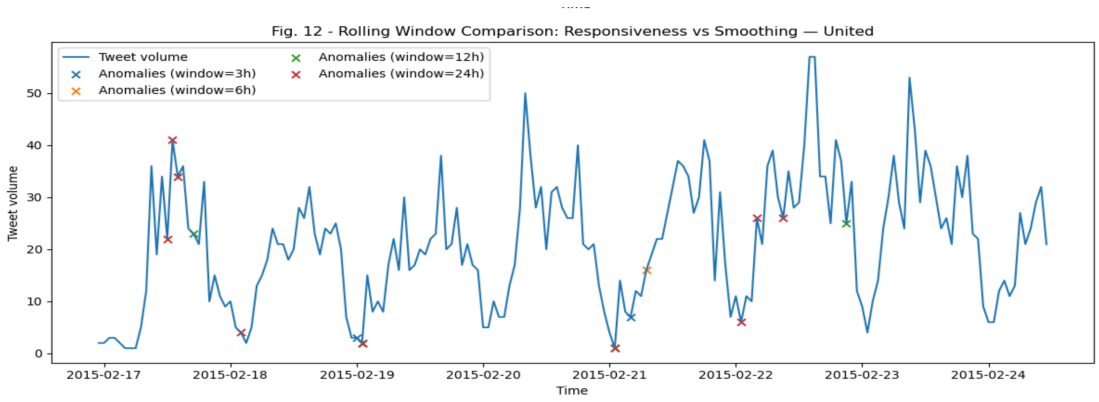
Feature design played a significant role. Sentiment-focused features aligned anomalies more closely with reputational risk, while volume-dominated features often flagged benign high-activity events. Incorporating engagement metrics improved differentiation between isolated complaints and widespread dissatisfaction, reinforcing the value of multi-dimensional feature design.

```
FEATURE SET: Sentiment-focused
- Isolation Forest sensitivity (contamination)
 contamination  anomaly_windows  anomaly_rate
          0.02                4      0.022222
          0.03                6      0.033333
          0.05                6      0.033333
          0.08               15      0.083333
          0.10               18      0.100000
Stable anomaly windows (>=60% settings): 6
```

Fig. 11- Sensitivity of Isolation Forest Anomaly Detection to Contamination Levels

11.2 Impact of Temporal Design Choices

The impact of temporal design choices was also investigated. The size of the rolling window significantly influenced detection granularity. Shorter windows enhanced responsiveness to rapid changes but introduced greater volatility, while longer windows reduced noise at the expense of delayed detection. Notably, despite these variations in timing and sensitivity, key anomaly periods were consistently identified across different window configurations. This suggests that the underlying behavioral patterns are resilient to reasonable variations in temporal design.



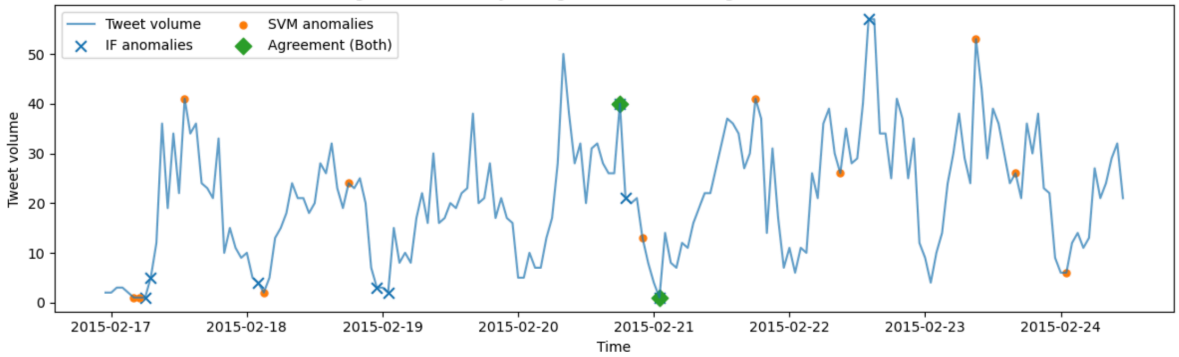
11.3 Reliability and Consistency of Early Warning Signals

Reliability was evaluated through temporal clustering and cross-model comparison. Detected anomalies tended to cluster around periods of heightened dissatisfaction and engagement rather than appearing randomly. Anomalies identified by both Isolation Forest and One-Class SVM were more likely to correspond to distinct behavioral shifts, while model disagreements primarily occurred in weaker or ambiguous cases.

It's important to note that not all anomalies resulted in sustained escalation, reflecting the probabilistic nature of early warning signals. The primary objective of detection is to highlight elevated risk, not to predict crises with absolute certainty. Overall, consistent patterns across time and models suggest that the framework effectively identifies recurring forms of early

reputational stress. When combined with human interpretation and ongoing recalibration, this framework serves as a reliable decision-support tool.

Figure 13 - Anomaly timing and cross-model agreement — United



12 Uncertainty, Risk, and Governance in Early Crisis Analytics

Early crisis detection based on social media data inevitably operates under conditions of uncertainty. Unlike retrospective analyses that can evaluate outcomes with hindsight, early warning systems aim to detect potential issues before their consequences become fully apparent. This section delves into the nature of uncertainty in early crisis analytics and the governance mechanisms necessary to manage associated risks responsibly.

12.1 Analytical Uncertainty and Risk Trade-Offs

Analytical uncertainty in social media-based crisis detection stems from various sources. User-generated content is often noisy, emotionally charged, and ambiguous. Negative sentiment doesn't always signify genuine dissatisfaction, and genuine dissatisfaction doesn't always escalate into a reputational crisis. Moreover, external factors like news coverage, weather events, or platform-specific trends can independently influence conversation patterns, unrelated to brand performance.

This uncertainty creates a fundamental trade-off between false positives and false negatives. Highly sensitive systems may detect emerging issues early but at the cost of generating alerts for non-escalating events. Conversely, conservative systems may miss early signals to reduce false alarms. The study's findings suggest that neither extreme is desirable. Instead, early warning systems should be designed to surface *risk indicators* rather than definitive judgments, allowing human decision-makers to assess context and significance.

Importantly, uncertainty shouldn't be viewed as a flaw to be eliminated but as an inherent feature of early-stage detection. The objective isn't perfect prediction but informed awareness. When framed in this manner, anomaly detection becomes a decision-support tool rather than an automated crisis classifier.

12.2 Governance, Ethics, and Accountability

Governance, ethics, and accountability are crucial aspects of managing associated risks responsibly in early crisis analytics. These elements ensure that the detection and response processes are transparent, accountable, and aligned with ethical principles.

Governance involves establishing clear frameworks and policies that govern the use of social media data for crisis detection. This includes defining the scope of data collection, ensuring data privacy and security, and establishing accountability mechanisms for those responsible for data handling and analysis.

Ethics play a vital role in maintaining the integrity and reliability of crisis detection systems. Ethical considerations include ensuring that data is used responsibly, avoiding bias and discrimination, and promoting transparency and accountability.

Accountability is essential for holding individuals and organizations accountable for their actions and decisions related to crisis detection. This includes establishing mechanisms for reporting and addressing concerns and ensuring that those responsible for crisis detection are held accountable for their performance.

Given the interpretive nature of early warning signals, governance frameworks are crucial in ensuring responsible use. Organizations must establish clear processes for reviewing anomaly alerts, identifying decision-makers with authority to act on them, and effectively communicating uncertainty internally. Without such structures, there's a risk of ignoring or over-interpreting analytics outputs.

Ethical considerations extend beyond data privacy to include accountability for decisions influenced by automated systems. While this study uses anonymized and publicly available data, the broader application of early warning systems raises concerns about reputational harm, misattribution of responsibility, and unfair escalation of issues. Transparency about model limitations and cautious language in reporting are therefore essential.

A governance-oriented approach emphasizes human oversight, documentation of decisions, and proportional responses to early signals. By embedding anomaly detection within structured organizational processes, firms can harness its benefits while mitigating ethical and reputational risks.

13 Organizational and Strategic Adoption Considerations

The effectiveness of early crisis analytics hinges not only on technical prowess but also on organizational readiness and strategic alignment. Even robust anomaly detection models may yield limited value if poorly integrated into decision-making processes or misunderstood by stakeholders. This section delves into the organizational prerequisites for effective adoption and sustained use of early warning systems.

13.1. Data Infrastructure and Analytical Capability

Successful implementation of early warning analytics demands a reliable data infrastructure and sufficient analytical expertise. Organizations must possess the capability to collect, process, and maintain high-quality social media and engagement data in near real time. Without this foundation, anomaly detection outputs may manifest as inconsistent, delayed, or difficult to interpret.

Analytical capability is equally crucial. Teams with experience in data-driven decision-making are better positioned to interpret probabilistic outputs, recognize uncertainty, and contextualize signals within broader operational knowledge. Conversely, organizations with lower analytical maturity may anticipate definitive predictions, leading to frustration or mistrust when models instead highlight elevated risk. Aligning model complexity with organizational capability is therefore paramount for adoption.

13.2. Process Integration and Operational Alignment

For early warning systems to deliver tangible value, they must be seamlessly integrated into existing crisis management, customer service, and communication workflows. Anomaly signals should trigger well-defined actions, such as internal reviews, targeted customer outreach, or monitoring escalation thresholds. Without this operational linkage, analytics outputs risk remaining isolated insights rather than actionable intelligence.
Strategic alignment is equally important. Senior management support ensures that early warning analytics are viewed as a decision-support tool rather than a purely technical experiment. Clear ownership of response protocols and accountability structures further strengthens the translation of insights into timely action.

13.3 Change Management and Organizational Trust

Adopting analytics-driven monitoring can raise concerns among employees about increased surveillance, workload, or erosion of professional judgment. Effective change management is therefore crucial. Organizations must communicate the purpose of early warning systems transparently, emphasizing their role in supporting human decision-making rather than replacing it.

Building trust also requires acknowledging model limitations. Presenting anomaly detection as a probabilistic risk indicator, rather than a definitive diagnosis, helps set realistic expectations and reduces resistance. Over time, consistent and interpretable signals can reinforce confidence and encourage broader organizational acceptance.

14 Strategic Value of Early Warning Systems in Brand Risk Management

Beyond operational benefits, early warning systems offer strategic value by supporting long-term brand resilience. In industries such as aviation, where customer trust and reliability are pivotal to brand equity, the ability to detect and address emerging issues early can shape competitive positioning.

Early detection enables organizations to respond promptly before negative narratives gain traction. Even when signals are ambiguous, timely investigation and engagement demonstrate responsiveness and concern, mitigating potential reputational damage. Over time, this fosters stronger customer relationships and enhances brand credibility.

Delayed response also incurs significant costs. As dissatisfaction spreads, recovery efforts become more expensive and less effective. Proactive intervention, guided by early warning signals and handled judiciously, can reduce escalation costs and limit long-term impact.

From a strategic standpoint, early warning systems should be regarded as investments in brand risk management rather than merely technical tools. By integrating early detection capabilities into decision-making processes, organizations can bolster their ability to navigate uncertainty and safeguard long-term brand equity in the ever-changing digital landscape.

15 Results Interpretation & Analytical Insights

This section moves beyond model outputs to interpret what the detected anomalies mean in practice. Rather than viewing anomalies as abstract statistical results, the analysis links them to observable changes in how users discuss airline brands on Twitter. This helps assess whether the detected patterns genuinely reflect early warning signs of emerging brand-related issues.

15.1 Characteristics of Detected Anomaly Periods

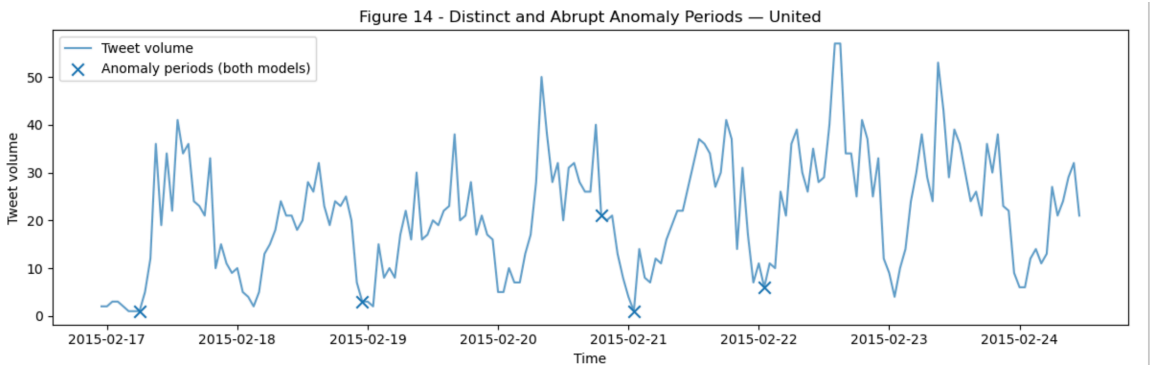
The anomaly detection models identified a small number of distinct time periods that differed noticeably from normal social media behaviour. These periods were not defined by long-term trends, but by **sudden and unexpected changes**, supporting the idea that early-stage crises tend to emerge abruptly rather than gradually.

A common feature across these anomaly periods was a sharp rise in tweet activity over short time windows. However, unlike routine increases in activity, such as those driven by promotions or general travel discussions, these spikes were dominated by negative or complaint-focused tweets. This suggests that the anomalies were not simply moments of high attention, but periods where dissatisfaction became unusually concentrated.

Another important characteristic was the speed at which sentiment changed. In anomaly periods, the proportion of negative sentiment increased rapidly relative to recent baseline levels. This acceleration is particularly important, as it indicates growing frustration among users rather than a steady background level of complaints. Engagement patterns further distinguished anomaly periods from routine activity. Negative tweets posted during these windows often received more retweets and replies than usual, increasing their visibility and potential impact. This amplification suggests that dissatisfaction was not only increasing but also spreading more widely across the network.

When anomalies were detected by both Isolation Forest and One-Class SVM, they were more likely to show clear thematic consistency. Multiple users tended to reference similar issues, such as delays or poor customer service, within the same time frame. This consistency

strengthens the interpretation of these periods as meaningful early warning signals rather than isolated or random outliers.

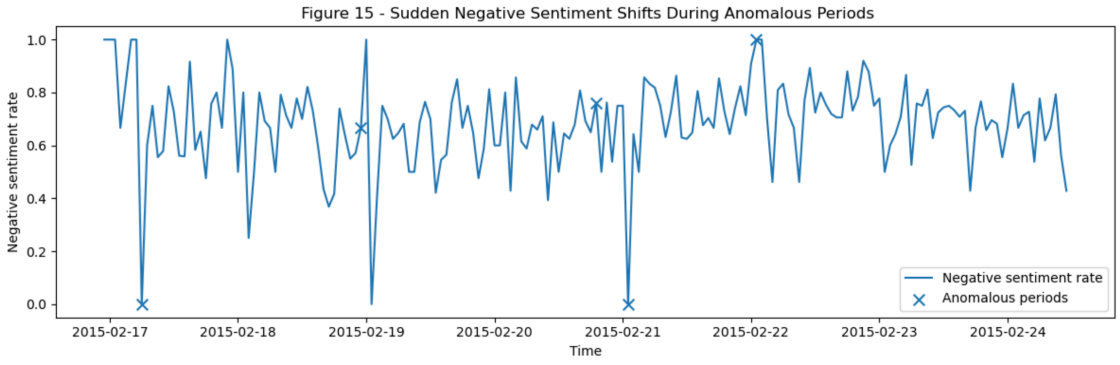


15.2 Comparison of Anomalous and Baseline Periods

Comparing anomalous periods with baseline periods reveals clear differences in how conversations evolve. During baseline periods, tweet volumes are relatively stable, sentiment is mixed, and discussions cover a wide range of topics. These patterns reflect routine social media activity related to travel and customer experiences.

In contrast, anomalous periods are more concentrated in both time and content. Tweet volume increases sharply within short windows, exceeding what would normally be expected based on recent behaviour. While these increases are often brief, they are significant enough to indicate early escalation rather than routine variation. Sentiment patterns also differ noticeably. Baseline periods typically show gradual shifts in sentiment, whereas anomalous periods are marked by sudden increases in negative sentiment without a corresponding rise in positive sentiment. This imbalance points to reputational stress rather than normal conversational dynamics. Engagement behaviour further highlights the distinction. During baseline periods, engagement tends to be spread across tweets regardless of sentiment. During anomalous periods, negative tweets attract a much larger share of retweets and replies, suggesting that critical content is gaining disproportionate attention.

Overall, these comparisons show that anomalous periods represent a different behavioural state from everyday social media activity, reinforcing their value as early indicators of potential brand crises.

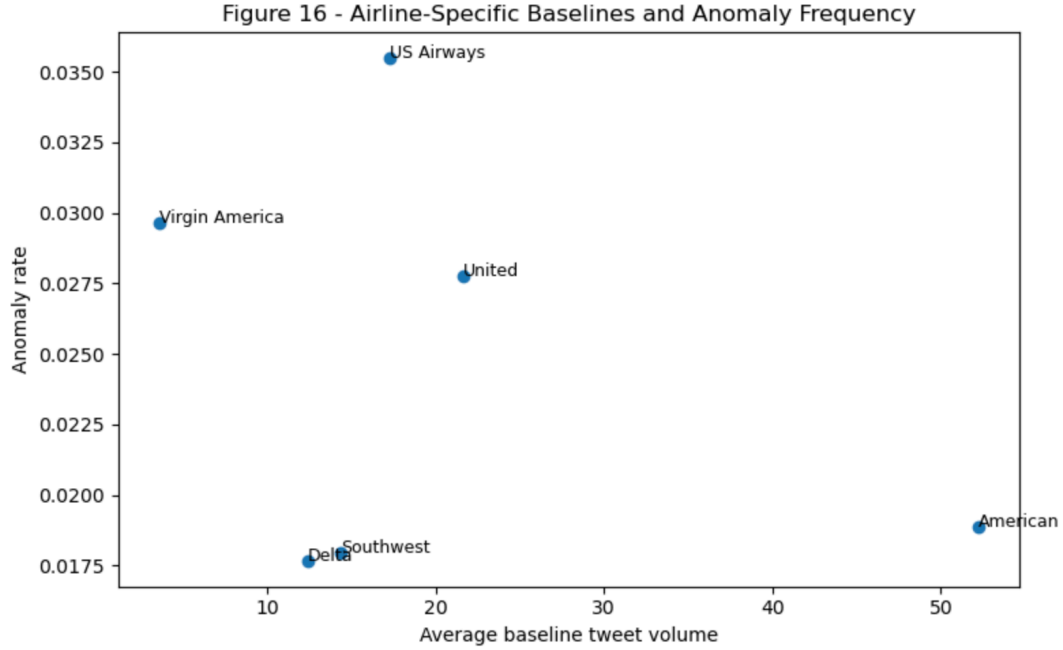


15.3 Airline-Level Differences in Anomaly Patterns

The analysis also shows that anomaly patterns vary across airlines, reflecting differences in brand visibility, customer expectations, and typical social media activity levels. Some airlines experience frequent but relatively mild anomalies, while others show fewer anomalies that are more intense when they occur. Airlines with consistently high tweet volumes tend to generate more anomalies in absolute terms, but these often represent smaller deviations from their normal activity levels. In contrast, airlines with lower baseline activity show fewer anomaly periods, but when anomalies do occur, they often involve sharper increases in negative sentiment and more concentrated complaints.

Differences are also evident in what drives the anomalies. For some airlines, anomalies are mainly associated with sudden increases in tweet volume, suggesting heightened attention linked to operational issues. For others, anomalies are driven more by sentiment changes, indicating rapid shifts in customer attitudes even without large volume spikes. Engagement patterns further distinguish airlines. Certain brands see negative tweets spread more widely during anomaly periods, increasing reputational risk. These differences highlight the importance of considering airline-specific baselines when interpreting early warning signals.

Figure 16 - Airline-Specific Baselines and Anomaly Frequency



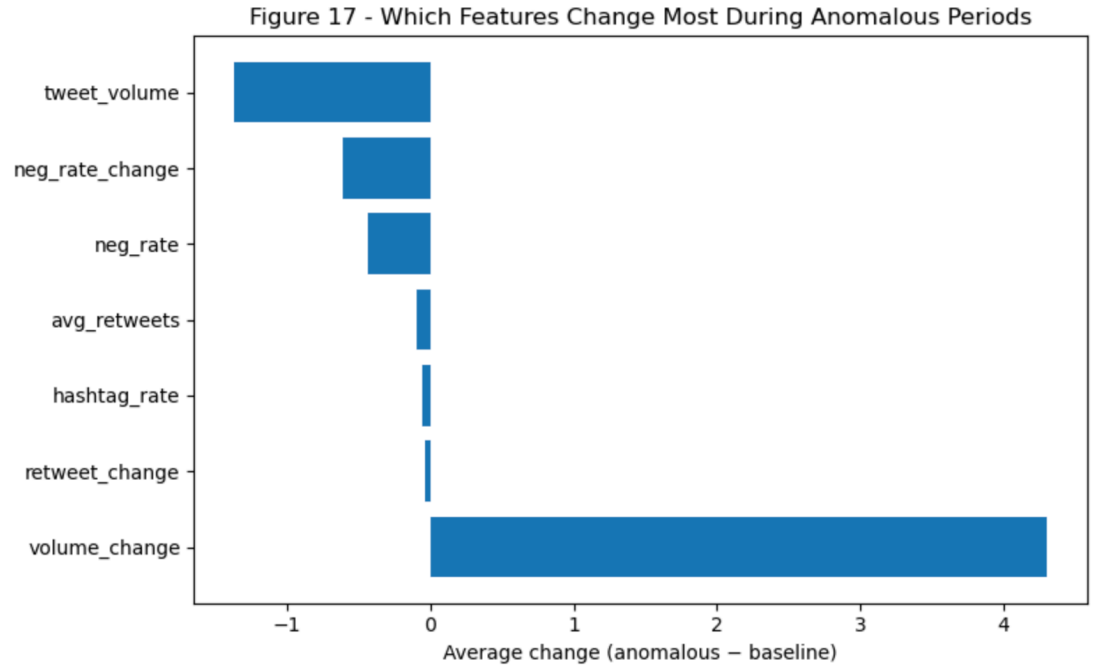
15.4 Feature Contribution Patterns in Early Crisis Signals

Examining which features contribute most to anomaly detection provides insight into how early crisis signals form. Across both models, changes in negative sentiment—particularly rapid increases over short periods—consistently emerge as the strongest indicators.
Tweet volume also plays an important role, but on its own it is rarely sufficient. High activity levels only become meaningful early warning signals when they coincide with worsening sentiment. This finding reinforces the limitations of monitoring volume alone, which remains common in many social listening tools. Engagement-related features, such as retweet activity,

add an important layer of interpretation. Anomalies associated with higher engagement are more likely to represent emerging issues, as they indicate that negative content is being shared rather than remaining isolated.

Hashtags and complaint-related keywords contribute in more targeted ways, often helping to clarify the nature of the issue rather than acting as primary detection signals. Together, these findings suggest that early crisis detection relies on a combination of signals rather than any single metric.

Figure 17 - Which Features Change Most During Anomalous Periods



16 Managerial & Practical Implications

This section discusses what the findings mean for practitioners, particularly those responsible for managing airline brand reputation and customer communications.

16.1 Implications for Airline Brand and Reputation Managers

The results underline the value of moving from reactive crisis management to a more proactive approach. By identifying unusual patterns early, brand teams can become aware of potential issues before they escalate into highly visible crises.

Rather than responding to individual complaints in isolation, anomaly detection allows managers to focus on periods where dissatisfaction is increasing unusually fast. This helps prioritise attention and resources toward issues that are more likely to spread and cause reputational damage. The findings also show that the speed and spread of negative sentiment matter as much as its volume. Monitoring how quickly negativity grows and how widely it is shared provides a clearer picture of escalation risk than counting complaints alone.

16.1 Integrating Anomaly Detection into Existing Social Listening Systems

Anomaly detection is best viewed as a complement to existing social listening tools rather than a replacement. It can act as an additional analytical layer that highlights unusual behavior requiring closer human review. In practice, anomaly alerts can prompt communication teams to investigate emerging issues, review recent tweets, and assess whether intervention is needed. Keeping humans in the loop helps ensure that statistical signals are interpreted within their proper context. Because the models are unsupervised, they are well suited to dynamic environments where new issues emerge unpredictably, and labelled training data are unavailable.

16.2 Decision-Making Under Uncertainty in Early Crisis Signals

Early warning systems inevitably involve uncertainty. Anomalies signal increased risk, not confirmed crises. The findings emphasise the importance of treating these signals as prompts for investigation rather than definitive conclusions. Managers should therefore approach anomaly alerts with caution, combining them with contextual knowledge and professional judgment. Clear communication about uncertainty can help prevent overreaction while still enabling timely responses. By framing anomaly detection as a decision-support tool rather than an automated decision-maker, organisations can use it responsibly and effectively.

17 Conclusion

This dissertation explored whether anomaly detection techniques can be used to identify early warning signs of brand-related crises in social media data from the airline industry. Using the Twitter US Airlines Sentiment Dataset, unsupervised machine learning models were applied to detect unusual patterns in sentiment, volume, and engagement.

The results show that early crisis signals are less about absolute levels of negativity and more about sudden departures from normal behaviour. Rapid increases in negative sentiment, especially when combined with elevated engagement, emerged as particularly strong indicators of potential escalation.

From a theoretical perspective, the study supports the view that brand crises often develop gradually through accumulating signals rather than appearing as single, clearly defined events. Methodologically, it demonstrates that interpretable, unsupervised models such as Isolation Forest and One-Class SVM can provide meaningful early warning insights without relying on labelled crisis data.

Practically, the findings suggest that anomaly detection can enhance existing social listening practices by helping organisations identify emerging risks sooner and respond more strategically. While the study has limitations, including its reliance on a single historical dataset, it provides a foundation for future research into proactive, data-driven crisis management.

Overall, the research shows that anomaly detection offers a promising approach to early brand crisis identification, supporting more timely and informed decision-making in high-risk, fast-moving social media environments.

18 Limitations

Publicly available social media datasets may not fully encompass all brands, platforms, or regional behaviors. External factors like real-world events, market trends, or media coverage can influence social media activity but may not be reflected in the data. Additionally, complex machine-learning models, such as LSTM autoencoders, can be challenging for non-technical stakeholders to interpret. Furthermore, findings may not generalize to all industries or types of social media crises.

19 Ethical Considerations

Dataset terms, conditions, and permissions

The datasets used in this project were sourced from publicly available repositories that offer pre-collected and anonymized Twitter data for research and educational purposes. Data collection and redistribution were conducted in accordance with the original platform terms of service and the licensing conditions specified by the dataset providers. The project does not involve direct access to the Twitter API or the collection of new data. Furthermore, no attempt is made to circumvent platform restrictions or rehydrate deleted content. The use of the data is strictly limited to non-commercial academic research, as per the stated permissions of the dataset sources. The project does not attempt to reproduce or republish raw content, and outputs are restricted to aggregated statistics and derived features. This approach ensures compliance with both platform policies and institutional research guidelines.

Timestamp handling and aggregation consistency

Although timestamp information is necessary for temporal analysis and anomaly detection, it is used solely to facilitate aggregation into fixed hourly time windows. This approach avoids tracking individual user behavior. All timestamps are converted to a standard datetime format and immediately aggregated. Subsequently, individual-level temporal information is discarded. This ensures consistency between timestamp handling and the project's focus on aggregate behavioral patterns rather than individual activity. As a result, no fine-grained temporal identifiers are retained that could be used to infer posting habits or re-identify users. This approach aligns with temporal modeling requirements and ethical principles of data minimization and purpose limitation. It ensures that time-based analysis supports early warning detection without introducing unnecessary privacy risks.

20 Associated Risks

Technical and methodological risks are acknowledged in this project, although it is considered low risk. Poor-quality or incomplete social media data may affect model performance, but this is mitigated through systematic data cleaning, timestamp validation, and hourly aggregation to reduce noise from individual tweets. To reduce the risk of model overfitting and limited generalisability, two conceptually distinct unsupervised anomaly detection techniques—Isolation Forest and One-Class SVM—are employed, and the anomaly periods identified by both models are prioritized. Additionally, sensitivity analysis is conducted by varying key hyperparameters and rolling-window sizes to assess the stability of detected patterns and ensure that results are not driven by arbitrary modeling choices.

A separate set of risks relates to interpretation and potential reputational implications of model outputs, particularly if results are misunderstood or treated as definitive indicators of crises. This risk is mitigated by explicitly framing anomalies as early-warning signals rather than confirmed events. Model outputs are supported by transparent feature analysis and manual qualitative inspection, and anomaly periods are interpreted only when supported by multiple validation layers, including cross-model agreement, rule-based thresholds, and contextual coherence in tweet content. By avoiding automated labeling and emphasizing human oversight, the project reduces the risk of misleading conclusions while maintaining responsible and proportionate use of unsupervised machine-learning methods.

21 References

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications, 60*, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv*. https://arxiv.org/abs/1901.0340

Deloitte. (2023). *2023 Global marketing trends: Accelerating innovation in the digital era*. Deloitte Insights. https://www2.deloitte.com

Keller, N., & Lee, S. Y. (2022). Crisis communication in the era of social-media analytics: Challenges and opportunities for early-warning systems. *Public Relations Review, 48*(1), 102–118. https://doi.org/10.1016/j.pubrev.2021.102118

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Yu, S., Liu, J., & Wu, S. (2016). Anomaly detection for social media data: An overview. *IEEE Access, 4*, 5857–5877. https://doi.org/10.1109/ACCESS.2016.2592958

Zimbra, D., Abbasi, A., & Chen, H. (2018). A cyber-archaeology approach to social media analytics: Investigating crisis events, trends, and interactions. *Journal of Management Information Systems, 35*(2), 450–489. https://doi.org/10.1080/07421222.2018.1451961

Appendix

Github Link – https://github.com/Pavan8765/Dissertation

One drive Link - https://universityofexeteruk-my.sharepoint.com/my?id=%2Fpersonal%2Fpa468%5Fexeter%5Fac%5Fuk%2FDocuments%2FDissertation&viewid=aa50d56e%2Dba56%2D47ec%2D9b1f%2Dd70ec6f46f09