Predictive and Comparative Analysis of Social Media Platform Performance for SMEs in Ireland and Nigeria

Olamide Abioro
Student ID: x23428902

January 28, 2026

Abstract

Small and medium sized enterprises increasingly rely on social media platforms to reach customers, yet it is still unclear which platforms and content strategies deliver the strongest engagement across different markets. This project builds a comparative prediction framework for Irish and Nigerian SMEs on Instagram, TikTok and Facebook. Using Apify based scrapers and a shared PostgreSQL database, a final sample of 40,217 public posts was collected from roughly 750 SME accounts across beauty, fashion and retail sectors.

After cleaning, cohort sizes ranged from roughly 650 to 2,700 posts per country sector platform combination. For each post, an engagement rate per 1,000 followers was constructed, and Random Forest and XGBoost classifiers were trained to predict whether a post would fall into the top quartile of engagement within its cohort, using temporal splits for validation. On held out test sets, the best cohorts achieved area under the ROC curve values between 0.80 and 0.93. Nigerian retail on Facebook and Nigerian fashion on TikTok achieved top decile lift values above 7, meaning the models captured more than seven times as many high engagement posts as a random selection of the same size.

The results show consistent gains from data driven posting strategies, with clear differences across platforms and sectors. Fashion and beauty SMEs benefit most from short video and mid length captions on Instagram and TikTok, while retail SMEs obtain comparatively stronger returns on Facebook. The framework is reusable for other markets and can support practical decision making for SME social media planning.

1 Introduction

Social media platforms such as Instagram, TikTok and Facebook have become central marketing channels for small and medium sized enterprises (SMEs). They allow businesses to communicate directly with customers, promote products and services, and cultivate brand communities at a comparatively low financial cost (Bruce et al. 2023, Lestari et al. 2024, Sharabati et al. 2024). The shift from traditional advertising toward digital and social media marketing has opened opportunities for SMEs to compete with larger brands through creative, targeted, and persistent content strategies (Chatterjee & Kar 2020).

However, the social media environment in which SMEs operate is strongly shaped by local digital infrastructure and user behavior. Ireland, a high income and digitally mature economy, reported internet penetration of 99.2 percent in 2025, with 77.8 percent of the population using social media (Kemp 2025*a*). Nigeria, classified as a lower middle income country, had an internet penetration rate of 45.4 percent and a social media usage rate of 16.4 percent in the same year (Kemp 2025*b*). These figures highlight a substantial digital divide between the two countries. They suggest that engagement dynamics that hold in Ireland may not simply carry over to Nigeria, where connectivity constraints, device costs, and platform preferences differ.

The academic literature has documented many benefits of social media adoption for SMEs, including improved customer communication, broader market reach, and stronger brand awareness (Durkin et al. 2013). Yet, several gaps remain. First, many studies rely on survey or interview data rather than large scale behavioral traces, which limits the ability to identify fine grained engagement patterns. Second, existing empirical work often focuses on a single platform or a single country, so there is little comparative evidence across platforms and national contexts. Third, only a small subset of studies

apply machine learning methods to predict which posts will achieve high engagement, and even fewer do so for SMEs rather than large brands or influencers.

This thesis addresses these gaps by posing three research questions:

1. To what extent can machine learning models predict high engagement posts for SMEs on Instagram, TikTok and Facebook
2. Which features relating to content, timing and account history are most predictive of engagement within each platform country sector cohort
3. How do effective posting strategies differ between Ireland and Nigeria and between sectors such as fashion, beauty and retail

To answer these questions, the study constructs a multi platform, cross country and sector specific pipeline. Social media posts from SME accounts in Ireland and Nigeria are scraped using Apify actors and stored in a PostgreSQL database. After cleaning and filtering, engagement based labels are defined using log transformed and account demeaned engagement rates, with high performing posts identified as those in the top quartile or decile of their cohort. A rich set of textual, temporal, historical and account level features is engineered. Random Forest and XGBoost classifiers are trained with temporal validation to predict whether a post will fall into the high engagement group. Model performance is assessed using Area Under the ROC Curve, Average Precision and top decile lift.

The contribution of this work is threefold. First, it delivers a reproducible multi platform analytics pipeline that can be used to model social media engagement for SMEs in both a developed and an emerging market. Second, it provides empirical evidence on cross platform and cross country differences in engagement dynamics, highlighting where predictive models perform best and which features matter most. Third, it translates these findings into practical guidance on content and timing strategies that SME managers can apply when planning social media campaigns. The remainder of this thesis is structured as follows.

- **Chapter 2: Related Work**
- **Chapter 3: Research Methodology**
- **Chapter 4: Design Specification**
- **Chapter 5: Implementation**
- **Chapter 6: Evaluation and Discussion**
- **Chapter 7: Conclusion and Future Work**

Figure 1 presents a conceptual modelling framework for SME social media performance. Public posts are collected via Apify and stored in PostgreSQL. Cohorts are defined by country, sector and platform. Features are engineered at post and account level, including timing, content and follower context. Within each cohort, temporal splits create training, validation and test sets. Random Forest and XGBoost models predict whether a post will fall into the top engagement quartile, and evaluation metrics such as AUC and top decile lift are compared across platforms and sectors to derive SME recommendations.

2 Related Work

This section reviews recent literature on social media use by small and medium sized enterprises and on the application of machine learning and predictive analytics to social media engagement. The review

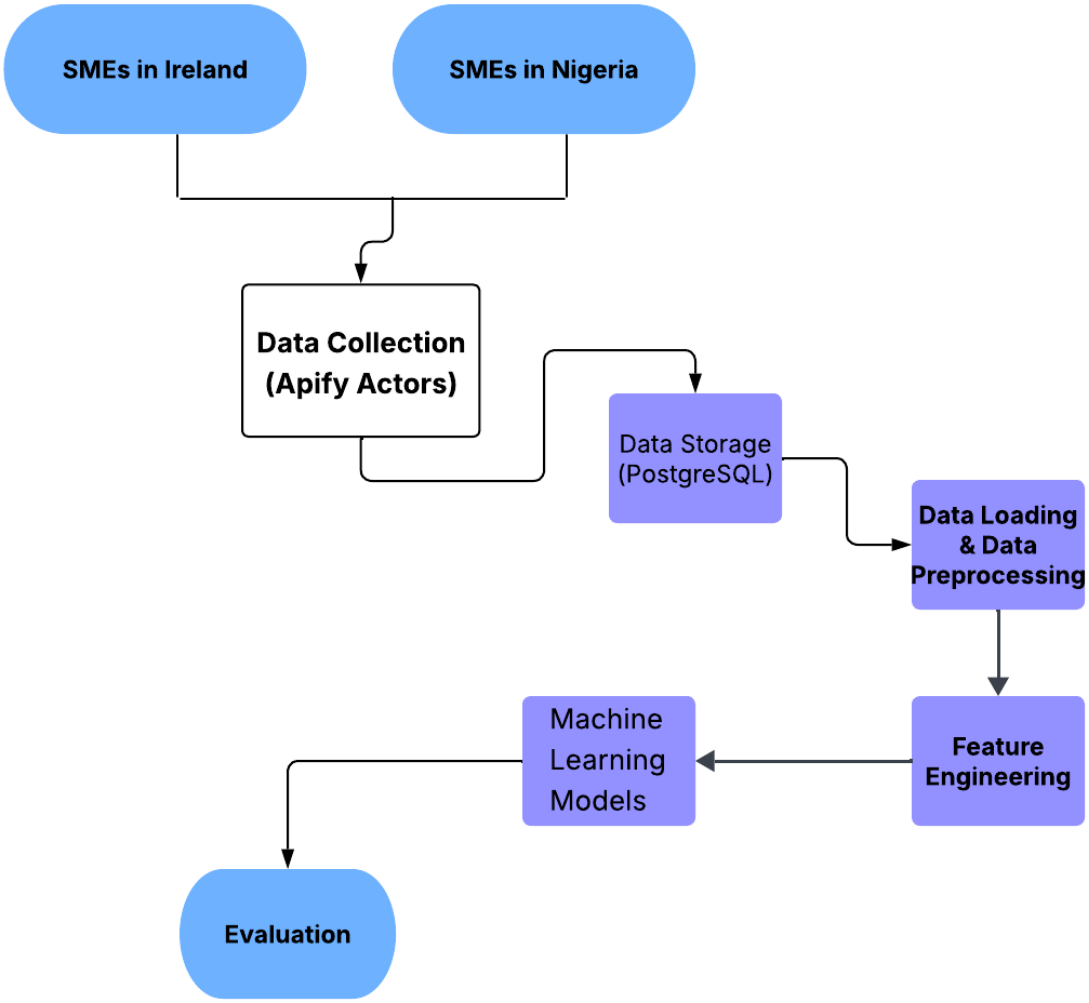


Figure 1: Conceptual modelling framework for SME social media performance.

is structured into two strands. Section 2.1 considers empirical research on social media marketing and performance among SMEs. Section 2.2 examines studies that employ machine learning or predictive methods to model social media engagement or related business outcomes. The section concludes with a synthesis of the main gaps that motivate the present study.

2.1 Social media marketing and SME performance

Recent work has reinforced the view that social media marketing is closely linked to SME performance, especially in emerging markets. Palladan *et al.* (2023) used a technology organisation environment framework to examine Nigerian SMEs and found that technological, organisational and environmental factors significantly influence social media adoption, while adoption itself has a positive effect on performance. They also reported that business strategy moderates the adoption performance relationship, which suggests that social media is most beneficial when integrated into a coherent strategic orientation rather than treated as an isolated activity.

Fu *et al.* (2024) extended this line of inquiry with a large scale survey of 508 Indonesian SMEs. Their study applied the technology organisation environment framework specifically to social media marketing adoption and SME performance. The results showed that organisational drivers such as top management support and employee capability and environmental pressures such as customer and competitor expectations positively shape adoption, whereas perceived technological barriers inhibit

it. Importantly, they found that adopting social media marketing is positively associated with SME performance, which provides strong quantitative support for the strategic role of social media in small firm growth.

Focusing directly on the link between social media activity and financial outcomes, Matosas López (2024) analysed an eight year panel of 100 SMEs to test whether customer engagement metrics on social media can be used to predict operating revenue. Using ARIMA models, the study found that customer engagement on social media does not predict revenue when social media is managed in house. However, engagement becomes a reliable predictor of revenue when social media management is outsourced to specialised providers and is combined with other digital marketing tools. This result highlights that the value of engagement signals depends on the broader marketing capabilities and organisational arrangements of the SME.

Several recent studies have examined social media marketing strategies and outcomes in specific sectors and national contexts. Casado Jiménez (2025) conducted a content analysis of Instagram and TikTok use by Spanish horticultural and fruit companies. The study found that interactive content that encourages two way communication and authenticity helps to build trust and engagement with digital native audiences. Haque *et al.* (2024) compared Instagram and TikTok as marketing channels and concluded that Instagram is particularly suited to visual storytelling and influencer based branding, while TikTok is more effective for viral, trend driven content and faster engagement cycles. These findings underscore that platform specific characteristics shape how SMEs should design their content strategies.

At a more synthetic level, Prasetya *et al.* (2025) carried out a systematic review of social media marketing studies published between 2020 and 2025. They identified influencer marketing, user engagement and the use of artificial intelligence and big data analytics as dominant themes in recent work. The review emphasises that authenticity, personalised content and long term relationship building are crucial for sustaining engagement and that emerging technologies such as live shopping and augmented reality are changing the nature of interaction between brands and consumers.

Taken together, these recent studies provide strong evidence that social media marketing can enhance SME performance and that the effectiveness of social media strategies is conditioned by organisational capabilities, sector characteristics and platform affordances. However, most of this work remains at the level of surveys, case studies or aggregate correlation. There is relatively little use of fine grained post level behavioural data to model engagement outcomes for SMEs, and there are few comparative analyses that examine differences between a highly connected economy such as Ireland and an emerging digital context such as Nigeria.

2.2 Machine learning and predictive analytics for social media engagement

In parallel, a growing literature applies machine learning and predictive analytics to social media data, although not always with a focus on SMEs. Kedi *et al.* (2024) reviewed machine learning applications for optimising SME social media marketing campaigns. They discussed how supervised learning models can support content recommendation, engagement prediction and audience segmentation and highlighted practical challenges such as data quality, lack of expertise and ethical considerations. Their work is primarily conceptual but concludes that machine learning has substantial potential to improve SME campaign efficiency through personalisation and real time analytics.

Several empirical studies have built predictive models on social media metrics to estimate business outcomes. Matosas López (2024), discussed above, used time series models to forecast SME revenue from social media customer engagement, showing that predictive power emerges under certain organisational conditions. Sanches and Ramos (2025) introduced a new compound metric, the loyalty rate, which combines interaction and virality rates on Instagram. In a case study of a digital agency, they used regression analysis to examine the relationship between loyalty rate and company turnover. They found a moderate positive correlation between loyalty rate and monthly revenue, although it did not reach conventional levels of statistical significance, and observed that follower count exhibited a stronger association with turnover. This work illustrates how engagement features can be linked to financial outcomes and motivates further research into richer predictive models.

Other studies focus more directly on engagement prediction. Santhiya and Kannan (2024) compared linear regression and Random Forest models for predicting the reach of Instagram Reels using features such as posting time, follower count and hashtag usage. Their results showed that Random

Forest achieved lower prediction error and identified posting time and follower base as influential predictors, reinforcing the suitability of tree based ensemble methods for social media data. Research on advertisement classification has also applied machine learning to distinguish between high performing and low performing social media ads, typically reporting that ensemble models outperform simpler baselines in ranking ads by expected response.

On the methodological side, the broader machine learning literature provides guidance on model choice and evaluation. Tree based ensembles such as Random Forests and gradient boosted trees have been shown to perform strongly on tabular marketing and customer data, and they remain standard baselines in recent churn prediction benchmarks (Bogaert & Delaere 2023, Plaatsman et al. 2025). Methodological work also highlights the importance of using validation schemes that respect temporal order in time dependent applications, together with ranking oriented metrics such as ROC AUC and lift for imbalanced classification problems (Bogaert & Delaere 2023, Plaatsman et al. 2025). In the churn prediction literature specifically, lift based measures, and in particular top decile lift, are widely used to quantify the practical value of a model for targeting scarce positive cases in direct marketing campaigns (Bogaert & Delaere 2023).

Although these works demonstrate that machine learning can effectively model social media engagement and related business outcomes, several limitations from the perspective of this thesis remain. First, many studies focus on single platforms or specific content types, such as Instagram Reels or Facebook ads, rather than comparing across multiple platforms in a unified framework. Second, most studies do not concentrate on SMEs as a distinct population and instead use data from large brands, influencers or mixed sets of accounts. Third, cross country comparisons are rare, and there is little work that explicitly contrasts engagement dynamics in developed and emerging economies using shared models and metrics.

2.3 Summary and research gap

The recent literature, therefore, establishes three key points that are directly relevant to this thesis. First, the adoption and strategic use of social media marketing are positively associated with SME performance in a range of contexts, including Nigeria and other emerging economies (Palladan et al. 2023, Fu et al. 2024, Matosas-López 2024). Second, platform specific features and content strategies, particularly interactive, authentic and visually rich posts on services such as Instagram and TikTok, play an important role in driving engagement (Casado Jiménez 2025, Haque et al. 2024, Prasetya et al. 2025). Third, machine learning models, especially tree based ensembles, have shown promise for predicting social media engagement and for linking engagement metrics to financial outcomes (Santhiya & Kannan 2024, Kedi et al. 2024, Sanches & Ramos 2025, Matosas-López 2024).

At the same time, there is a clear gap at the intersection of these strands. Few studies combine large scale post level data from SMEs, modern classification models and a comparative design that spans both a developed digital economy and an emerging one. Existing work rarely models organic SME posts across multiple platforms simultaneously, and cross national comparisons are typically based on survey data rather than behavioural traces. This thesis addresses this gap by constructing a unified multi platform predictive framework that uses SME social media data from Ireland and Nigeria, applies ensemble classifiers to predict high engagement posts and systematically compares performance and feature importance across platforms, sectors and countries.

3 Research Methodology

This chapter describes the research design, data collection procedure, preprocessing and feature engineering pipeline, modelling strategy and evaluation methodology used in the study. The goal is to provide a transparent and reproducible account of how raw social media data from SME accounts in Ireland and Nigeria were transformed into modelling ready datasets and how the classification models were trained and assessed. The methodological choices are informed by best practices in applied predictive modelling (Johnson & Kuhn 2025) and by prior work on marketing analytics and class imbalance in customer centric applications (Benslama & Jallouli 2022, Rosário & Dias 2023, Singh 2024, Chen et al. 2024, Imani et al. 2025).

3.1 Research design

The study adopts a quantitative, supervised learning design with a comparative structure across platforms, countries and sectors. Social media posts constitute the basic unit of analysis. Each post is represented by a set of features derived from its content, timing and account history and is associated with a binary target label that indicates whether it is a high engagement post relative to similar posts.

Cohorts are defined by the combination of country (Ireland or Nigeria) and sector (fashion, beauty or retail) for different platforms (Instagram, TikTok or Facebook). In principle this yields up to six cohorts. For each cohort with sufficient data, separate classification models are trained and evaluated. This design supports two aims. First, it allows prediction models that are tailored to specific country sector contexts. Second, it allows systematic comparison of model performance and feature importance across cohorts.

The overall workflow follows a standard predictive analytics pipeline described by Kuhn & Johnson (2013). The main stages are data collection and storage, data cleaning and preprocessing, feature engineering and target construction, model training and hyperparameter tuning and final evaluation on held out test sets.

3.2 Data collection

3.2.1 SME sampling frame

The study focuses on SMEs operating in the fashion, beauty and retail sectors in Ireland and Nigeria for three main reasons. First, these sectors are visually rich and already highly active on Instagram, TikTok and Facebook, which makes them natural candidates for social commerce and engagement based optimisation. Second, both Ireland and Nigeria have a large number of independent boutiques, cosmetics brands and gift or lifestyle retailers that maintain public social media accounts, which ensures sufficient data volume and balanced cohorts across countries. Third, prior work on SME social media performance often concentrates on consumer facing, visually oriented industries, so focusing on these sectors allows this project to connect back to, and extend, the existing literature while still providing actionable guidance for real businesses. Candidate SMEs were identified through online business directories, platform search and existing curated lists. Only businesses with public profiles or pages on Instagram, TikTok or Facebook and with clearly identifiable brand accounts were included. This ensured that scraped posts corresponded to legitimate SME marketing content rather than to individual personal accounts.

3.2.2 Platform data acquisition

Data collection was carried out using Apify, a cloud based web scraping platform that provides configurable actors for major social media sites. Custom extraction scripts were written in Python to orchestrate the scraping process and to ingest the results into a PostgreSQL database. In particular, the scripts `ingest_apify_instagram.py`, `ingest_apify_tiktok.py` and `ingest_apify_facebook.py` were developed.

For each platform, the Apify actor was configured with a list of SME account handles and instructed to retrieve recent posts and associated engagement metrics. The scraped output was returned as structured JSON, which was then parsed by the ingestion scripts. The scripts normalised the JSON into relational tables and inserted records into the database, handling incremental updates so that repeated runs did not create duplicate entries.

The final collected dataset comprised:

- 19 269 Instagram posts,
- 12 286 TikTok posts,
- 8 662 Facebook posts.

The lower volume of Facebook posts reflects both the greater technical complexity of scraping Facebook and practical constraints related to Apify usage quotas. For Facebook, additional care was taken to retrieve posts from 2025 in order to align the time window with the other platforms.

3.2.3 Database schema

A PostgreSQL database was used as the central data store. The schema was designed around two main tables:

- `accounts`, which stores account level information such as an internal account identifier, platform, country, sector, username and the highest observed follower count,
- `posts`, which stores post level information such as post identifier, foreign key to `accounts`, posted_at timestamp, content type and raw engagement metrics (likes, comments, shares, saves, views) together with the follower count observed at or near the time of posting,

3.3 Ethical considerations and data governance

The data used in this project consists entirely of public posts and public metadata from business accounts. No personal profiles were scraped and no attempts were made to infer information about individual users. All data were stored in a secure PostgreSQL database on a password protected machine and only aggregated statistics are reported. From a regulatory perspective, the work aligns with the principles of the European General Data Protection Regulation and the Nigerian Data Protection Act, since no directly identifiable personal data are collected and the analysis is carried out for academic research purposes. The Apify scrapers were configured to respect each platform's terms of service, in particular by limiting request rates, avoiding automated interactions with private accounts and treating the data as read only observational logs rather than as a channel for direct marketing.

3.4 Data preprocessing

The next phase involved cleaning and filtering the raw data to produce consistent, analysable cohorts.

3.4.1 Account and post level filtering

To focus on genuine SME presences and to avoid instability in engagement rate calculations for very small denominators, accounts with fewer than 300 followers were removed from the dataset. This threshold is in line with common practice in social media analytics where extremely small accounts are excluded to reduce noise.

Posts were then filtered to remove records with missing or invalid key fields. In particular, posts with missing timestamps (represented as `Nan`) or missing follower counts were discarded. Platform specific schema inconsistencies were resolved, for example by mapping alternative content type labels to a common vocabulary (image, video, reel and so on).

3.4.2 Promotional content removal

The focus of the analysis is on regular organic content rather than on heavily promotional or incentive based posts. To this end, posts whose captions contained clear markers of giveaways, competitions or explicit discount campaigns were removed using regular expression based filters. Typical patterns included words and phrases such as *giveaway*, *win*, *flash sale*, *use code* and similar. These posts often have distinct engagement dynamics driven by prize incentives, which could distort models aimed at everyday content optimisation.

3.4.3 Cohort assignment

Each post inherits platform, country and sector attributes from its parent account. These attributes were used to assign posts to cohorts of the form (country, sector). For each cohort, basic descriptive statistics such as post count and average engagement rate were computed as a preliminary check that sufficient data existed to support model estimation.

3.5 Feature engineering

Feature engineering transforms cleaned post records into a set of numerical and categorical predictors suitable for machine learning models. This process was carried out primarily in Python using pandas for tabular manipulation and NumPy for numerical operations (McKinney 2010, Harris et al. 2020).

3.5.1 Base engagement variables

For each post i belonging to account a, a total interaction count E_i and a follower count F_i at the time of posting were defined. The exact construction of E_i depended on the platform:

- Instagram and Facebook:
$$E_i = \text{likes}_i + \text{comments}_i$$
- TikTok:
$$E_i = \alpha_1 \text{views}_i + \alpha_2 \text{likes}_i + \alpha_3 \text{comments}_i + \alpha_4 \text{share}_i$$

  For TikTok posts, the raw interaction volume is dominated by video views, while likes, comments and shares represent progressively higher intent. If views are treated on the same scale as likes, they swamp the signal from more meaningful interactions. To address this, a weighted engagement score was defined before normalising by followers.

  Let views_i, likes_i, comments_i and shares_i denote the interaction counts for post i, and let followers_i be the creator follower count at the time of posting. The TikTok engagement score is defined as

$$E_i^{\text{TikTok}} \;=\; 0.02 \cdot \text{views}_i \;+\; 1.0 \cdot \text{likes}_i \;+\; 2.0 \cdot \text{comments}_i \;+\; 2.0 \cdot \text{shares}_i.$$

  The weights reflect the relative effort and business value of each action. Views are extremely frequent and mostly passive, therefore they receive a very small coefficient. Likes are a low friction but intentional signal and receive a baseline weight of 1. Comments and shares are rarer and indicate stronger involvement, so they are given a weight of 2. This choice is consistent with common practice in social media analytics, where deeper actions such as commenting and sharing are treated as more valuable than views or reactions. The goal of this weighting is not to discover a unique true engagement function, but to define a reasonable operational proxy that distinguishes highly interactive posts from background behaviour. Because the models use percentile based labels defined within each cohort, they depend mainly on the ranking induced by this score rather than its absolute scale. A full grid based sensitivity analysis of alternative weight combinations was beyond the scope of this work. I therefore treat the chosen coefficients as a defensible, domain informed heuristic, and note that systematic testing of alternative TikTok weightings would be a useful extension in future work.

  The per follower engagement rate used for modelling is then

$$ER_i^{\text{TikTok}} \;=\; \frac{E_i^{\text{TikTok}}}{\text{followers}_i}.$$

  This definition is consistent with the rate based measures used for Instagram and Facebook, in the sense that all platforms ultimately use a per follower engagement rate as the target. The difference is that for TikTok, the raw engagement score is first reweighted to downscale passive views and emphasise active interactions.

  The specific coefficients were chosen using exploratory analysis of the scraped TikTok data. Typical view to like ratios were in the tens to one range, so a weight of 0.02 for views makes one like roughly comparable to several tens of views. Doubling the weight for comments and shares aligns with their higher intent nature and is consistent with prior work that treats these actions as stronger indicators of user interest than likes. This hand tuned weighting scheme avoids the need for an additional learning step, while still capturing the relative importance of different interaction types on TikTok.

The basic engagement rate for post i was then defined as

$$ER_i = \frac{E_i}{F_i}. \tag{1}$$

This normalises interaction by audience size and allows comparison across accounts with different follower counts.

3.5.2 Log transformed and demeaned engagement

Engagement rates are typically highly skewed. To stabilise variance and reduce the influence of extreme outliers, a log transformed engagement rate was computed. A small constant $\varepsilon > 0$ was added to avoid taking the logarithm of zero. For each post i,

$$\text{log_rate}_i = \ln(ER_i + \varepsilon). \tag{2}$$

To separate post level variation from structural differences between accounts, the log rate was then demeaned at account level. For each account a with n_a posts, the mean log engagement rate is

$$\overline{\text{log_rate}}_a = \frac{1}{n_a} \sum_{i \in a} \text{log_rate}_i, \tag{3}$$

and the account demeaned log engagement for post i is

$$\text{log_rate_demeaned}_i = \text{log_rate}_i - \overline{\text{log_rate}}_a. \tag{4}$$

Positive values indicate posts that perform above the typical baseline for that account, while negative values indicate under performance. This transformation encourages the models to capture post level choices that SMEs can control, such as content and timing, rather than static differences in follower base.

3.5.3 Textual features

From each caption, a set of simple textual features was derived. These included:

- caption length in characters and in words,
- number of hashtags,
- number of user mentions,
- binary indicators for the presence of call to action phrases such as *link in bio*, *shop now*, *tag a friend*.

These features capture basic aspects of how much information is conveyed, how discoverable the post is through hashtags and whether the SME explicitly prompts the user to act. More advanced text features such as sentiment scores were considered, but the final models focused on these simpler, more interpretable quantities.

3.5.4 Temporal features

The timestamp of each post was decomposed into several temporal features:

- hour of day (0 to 23),
- day of week (Monday to Sunday),
- a coarse period of day category (for example, morning, afternoon, evening, night).

These variables encode known effects of posting time on engagement, since user activity patterns and platform algorithms vary across the day and week. Following recommendations in Kuhn & Johnson (2013), time variables were treated as categorical where appropriate.

3.5.5 Historical and account features

To capture the recent history and typical behaviour of an account, additional features were engineered:

- time since previous post by the same account,
- rolling average of engagement rate over the last N posts (for example the last five posts),
- total number of posts by the account in a recent window (posts per week),
- maximum and current follower count for the account,
- overall historical engagement rate of the account.

Several features in the final models use historical information at the account level, for example posts per week, time since the previous post and rolling engagement summaries. These variables are useful because they capture posting rhythm and recent audience behaviour, but they must be constructed carefully in order to avoid temporal leakage. All such features are computed separately for each account after sorting its posts by timestamp in ascending order. Rolling windows are defined so that, for a given post at time t, only events at or before t within the window contribute to the feature value. For example, posts per week counts the number of posts from the same account in the preceding seven days, and time since previous post is computed as the difference in minutes between the current timestamp and the immediately earlier post timestamp. No information from future posts is used when constructing features. Temporal splits are then applied after feature creation. Each cohort is ordered by posted time, and the most recent fifteen percent of posts are reserved for testing, with the previous fifteen percent used for validation and the remainder for training. Since all features for a given post are based only on that account's history up to that time, and label thresholds are computed using training data only, the validation and test sets simulate a realistic forward looking prediction setting without leakage from the future.

3.6 Target construction

The prediction task is formulated as binary classification on the account demeaned log engagement. Within each cohort defined by country, platform and sector, posts are ranked by $\text{log_rate_demeaned}_i$. Two labelling schemes were considered:

- **Top quartile**: posts in the top 25% of the demeaned distribution are labelled as high engagement ($y_i = 1$) and the remainder as non high engagement ($y_i = 0$),
- **Top decile**: posts in the top 10% are labelled as high engagement and the remainder as non high engagement.

The main analysis uses the top quartile labelling, with the top decile scheme employed for sensitivity checks. Defining success in terms of cohort specific percentiles has two advantages. First, it normalises for differences in baseline engagement levels across countries, platforms and sectors, so that a high engagement post is always defined relative to its peers rather than by an absolute threshold. Second, it aligns with common practice in marketing analytics, where lift based evaluation compares the concentration of positive cases in the top ranked fraction of a scored list to the overall prevalence (Burez & Van den Poel 2009).

The classifiers do not predict a continuous engagement rate directly. Instead, they identify posts that belong to the high engagement tail of the distribution within each country sector and platform cohort. For Instagram and Facebook, positive labels correspond to posts in approximately the top quartile of engagement rate within the training data for that cohort. For TikTok, where engagement distributions are more heavy tailed, positives are defined as posts in the top decile. Practically, for each cohort I compute engagement rate as engagement divided by followers, sort only the training portion of the data by this rate, and take the seventy fifth percentile threshold (or the ninetieth percentile for TikTok). Validation and test labels are then defined by comparing their engagement rate against the training threshold. This prevents leakage from future data into the labelling step and keeps class balance stable. Across cohorts, the resulting positive class prevalence remains in a moderate range, so the classifiers avoid degenerate situations where almost all posts are either positive or negative. Using

percentile based thresholds rather than absolute engagement values also makes labels comparable across cohorts that differ greatly in scale, for example small Nigerian retail accounts versus large Irish fashion brands.

3.7 Model hyperparameters and tuning

Two tree based ensemble families are used throughout the project. Random Forest provides a robust non parametric baseline, and gradient boosted trees through XGBoost allow finer control of model capacity and class imbalance. Hyperparameters were tuned once on the Instagram Ireland beauty cohort, which is one of the largest and cleanest segments in the data, and then reused unchanged across all other cohorts and platforms.

The tuning process followed a simple but explicit protocol. First, a sensible range was chosen for each key hyperparameter based on prior work, for example number of trees in the range 300 to 1,200, maximum tree depth between 4 and 8, and learning rate between 0.03 and 0.1 for XGBoost. Second, combinations from this grid were evaluated using temporal validation on the Ireland beauty training and validation split, and the setting that maximised validation area under the ROC curve while keeping overfitting under control was selected. The final Random Forest configuration uses 700 trees, square root feature subsampling at each split, a minimum of 3 samples per leaf, and balanced class weights so that minority high engagement posts are not ignored. The XGBoost configuration uses 1,200 trees, a learning rate of 0.05, a maximum depth of 6, row subsampling of 0.9, column subsampling of 0.8, and a regularization term of 2. The scale pos weight parameter is set to the ratio of negative to positive examples in each training cohort, which compensates for the relatively small fraction of high engagement posts.

These hyperparameters are then applied unchanged to all other country sector and platform cohorts. This reuse supports comparability across models, and avoids overfitting each cohort through extensive bespoke tuning. Generalisability was checked empirically by confirming that the same configuration produced stable and high validation and test AUC values across multiple Instagram, TikTok and Facebook cohorts. Where a cohort was very small, the models were not tuned further and the results are reported with appropriate caution. Table 1 summarises the main hyperparameters.

Table 1: Summary of classifier hyperparameters

Model	Hyperparameter	Value or rule
Random Forest classifier	`n_estimators`	700
	`max_depth`	None (trees grown to purity)
	`min_samples_leaf`	3
	`max_features`	square root of feature count
	`class_weight`	`balanced`
	`random_state`	42
XGBoost classifier	`n_estimators`	1 200
	`learning_rate`	0.05
	`max_depth`	6
	`subsample`	0.9
	`colsample_bytree`	0.8
	`reg_lambda`	2.0
	`scale_pos_weight`	ratio of negative to positive cases
	`eval_metric`	AUC on validation set

4 Design Specification

This chapter specifies the overall architecture and framework that underpin the implementation described in the next chapter. It identifies the main components of the system, the data structures and flows that connect them, and the way in which the predictive models are integrated into this pipeline. It also provides a word based description of the Random Forest and XGBoost algorithms used in the study and outlines the key functional and non functional requirements that the design must satisfy.

4.1 Overall system architecture

The system is designed as a modular analytics pipeline that moves from raw social media data to decision relevant outputs. Conceptually, the architecture can be viewed as five connected layers:

1. A data acquisition layer that collects posts and account information from Instagram, TikTok and Facebook using Apify actors,
2. A storage layer that persists the collected data in a PostgreSQL relational database,
3. A feature and label construction layer that derives modelling features and engagement based targets from the raw tables,
4. A modelling layer that trains and evaluates classification models for each country platform sector cohort,
5. An analysis and reporting layer that aggregates model outputs into tables, figures and strategic insights.

Figure 2 illustrates the high level architecture and the main data flows between these layers.

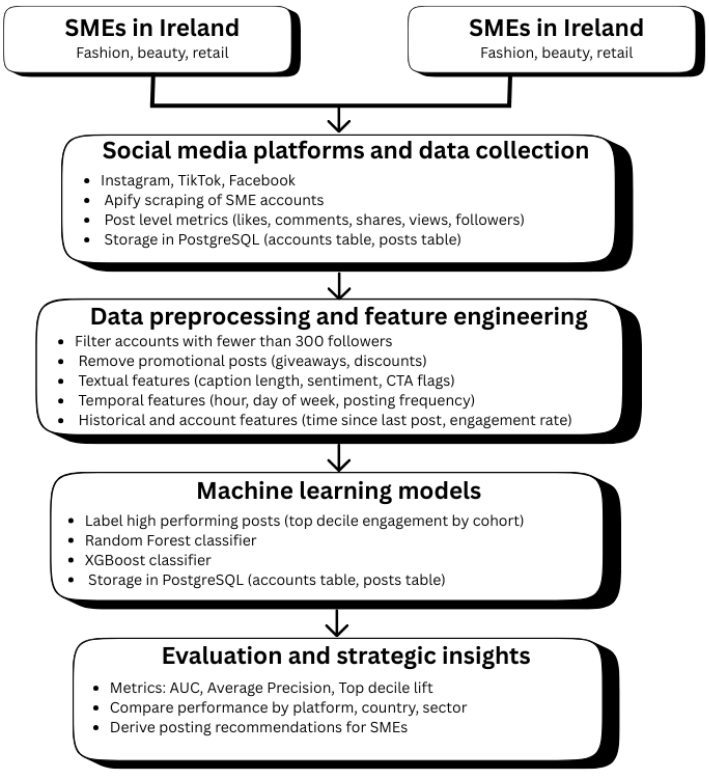


Figure 2: High level architecture of the SME social media analytics pipeline.

This layered design separates concerns. Data acquisition and persistence can evolve independently of feature engineering, while the modelling layer can be extended with additional algorithms without altering the upstream collection workflow.

4.2 Data and storage architecture

The storage layer is built around a PostgreSQL database that acts as the system of record for all scraped social media data. The schema follows the logical structure outlined in Section 3 and is summarised here from a design perspective.

- The `accounts` table holds stable information about each SME account. Key fields include a unique internal account identifier, platform, country, sector, username and the highest observed follower count. This table defines the set of entities whose behaviour is tracked over time.
- The `posts` table holds event like records for individual posts. Each row includes a post identifier, a foreign key to `accounts`, the posted_at timestamp, content type, caption text and platform specific engagement metrics together with the follower count at or near the time of posting.

This schema supports efficient joins between account and post data and allows cohort specific subsets to be extracted by filtering on platform, country and sector attributes. Using a relational database rather than flat files ensures that referential integrity is maintained and that complex filters and aggregations can be expressed in SQL and pushed down to the database engine.

4.3 Feature and label generation framework

The feature and label construction layer implements a deterministic transformation from cleaned database records to modelling ready tables. It can be thought of as a framework with a clear sequence of stages.

1. **Extraction stage.** For a given cohort specification, posts and their parent accounts are queried from PostgreSQL and loaded into pandas DataFrames. Only posts that satisfy basic quality filters are included.
2. **Base metric computation.** For each post, base engagement variables are calculated from likes, comments, shares, views and follower counts. The raw engagement rate is computed as the ratio of total interactions to follower count.
3. **Stability and normalisation.** The engagement rate is log transformed and demeaned at account level to create a stable, baselined measure of performance that isolates post level variation.
4. **Feature derivation.** Textual, temporal, historical and account level features are constructed as described in Section 3. Caption length, hashtag counts, posting hour and day of week, time since previous post and rolling engagement statistics are all included.
5. **Label assignment.** Within the cohort, posts are ranked by demeaned log engagement and assigned binary labels based on percentile thresholds (for example top quartile equals high engagement). The resulting table contains only feature columns and the target column.
6. **Materialisation.** The final feature label table for the cohort is written to disk in a serialised format or to a dedicated schema in PostgreSQL, so that the modelling layer can consume it without repeating preprocessing.

This framework ensures that all cohorts are generated using consistent logic. It also makes it straightforward to rerun the pipeline when new data are ingested or when feature definitions are refined.

4.4 Modelling framework and algorithm description

The modelling layer is responsible for learning mappings from the engineered features to the high engagement label for each cohort. The design centres on two ensemble tree based algorithms, Random Forest and XGBoost, chosen for their strong performance on structured data and their ability to handle nonlinear interactions (Breiman 2001, Chen & Guestrin 2016).

4.4.1 Random Forest

Random Forest can be described in words as follows. The algorithm builds many decision trees, each trained on a different bootstrap sample of the training data. At each split in a tree, only a random subset of features is considered. Individual trees are grown to sufficient depth so that they capture detailed patterns, but each tree on its own is noisy. To make a prediction for a new post, the Random

Forest sends the feature vector down every tree, collects the class prediction from each tree and then averages these predictions to obtain a final probability of high engagement.

This design reduces variance through averaging and makes the model more robust to outliers and noise than a single decision tree. The random selection of features at each split also decorrelates trees, which improves ensemble performance. In the context of this study, Random Forest is attractive because it can naturally handle mixed types of features, such as numerical follower counts and categorical posting hours, without extensive preprocessing.

4.4.2 XGBoost

XGBoost is a gradient boosting algorithm that also uses decision trees as base learners (Chen & Guestrin 2016). Unlike Random Forest, which builds trees independently and averages their predictions, XGBoost builds trees sequentially. The process starts with a simple model, typically a constant prediction. At each subsequent step, a new tree is trained to predict the residual errors of the current ensemble. In other words, each tree tries to correct the mistakes of the previous trees.

The contribution of the new tree is added to the ensemble with a learning rate that controls how strongly it influences the overall prediction. Regularisation terms and constraints on tree depth and structure are used to prevent overfitting. For classification, XGBoost outputs a probability of high engagement obtained by applying a logistic link function to the aggregated tree outputs.

From a design standpoint, XGBoost offers several advantages. It provides fine control over model complexity, can incorporate class weighting through parameters such as `scale_pos_weight` and includes efficient handling of sparse features. It is well suited to the moderately sized, tabular datasets used in this study and has been widely adopted in marketing and social media prediction tasks.

4.4.3 Cohort based model organisation

The modelling framework treats each cohort as a separate prediction problem. For each country sector combination with sufficient data, the system:

1. loads the corresponding feature label table,
2. splits the data into a chronological training and test set,
3. trains a Random Forest model with appropriate class weights,
4. trains an XGBoost model with tuned hyperparameters and positive class scaling,
5. stores the trained model objects and evaluation metrics for later comparison.

This organisation makes it easy to add or remove cohorts and to extend the framework to new countries or sectors. It also ensures that patterns specific to one cohort are not diluted by pooling with others.

4.5 Functional and non functional requirements

From a design perspective, the system is required to satisfy several functional requirements.

- It must ingest social media data from multiple platforms and normalise them into a common relational schema.
- It must construct features and labels in a consistent and reproducible way for any cohort specification.
- It must train classification models for each cohort and compute agreed evaluation metrics, including AUC, average precision and top decile lift.
- It must produce interpretable outputs, such as feature importance rankings and performance tables, that can support strategic recommendations for SMEs.

In addition, the architecture is designed to meet non functional requirements.

- **Reproducibility.** All transformations from raw Apify output to final model metrics are encoded in version controlled code and notebooks, which allows experiments to be rerun and audited.
- **Extensibility.** New platforms, sectors or countries can be incorporated by adding cohort configurations and, if necessary, additional Apify actors, without redesigning the core pipeline.
- **Scalability.** The use of a relational database and efficient Python libraries means that the pipeline can handle tens of thousands of posts per cohort with reasonable computational resources.
- **Transparency.** By using well understood algorithms such as Random Forest and XGBoost and by reporting feature importance, the system provides insight into how predictions are made, which is important for practitioner trust.

These requirements ensure that the design is not only technically sound but also aligned with the practical goal of delivering actionable insights on social media strategy for SMEs in Ireland and Nigeria.

5 Implementation

This chapter describes the final implementation of the social media analytics pipeline. Earlier exploratory versions of the code and provisional experiments are not discussed. The focus is on the stable set of scripts, notebooks and artefacts that produced the results reported in the evaluation chapter, and on the tools and languages used to generate these outputs.

5.1 Technical environment

All components of the system were implemented in Python 3. The main data analysis stack consisted of pandas for tabular data manipulation (McKinney 2010), NumPy for numerical operations (Harris et al. 2020), scikit learn for model fitting and evaluation (Pedregosa et al. 2011) and the XGBoost Python library for gradient boosted tree models (Chen & Guestrin 2016).

PostgreSQL served as the relational database management system for persistent storage of accounts and posts (PostgreSQL Global Development Group 2023). Data extraction and cleaning were performed in Python scripts and Jupyter Notebooks, which provided an interactive environment for combining code, narrative and visual output (Kluyver et al. 2016). Plots such as ROC curves, precision recall curves and feature importance charts were created using Matplotlib (Hunter 2007).

The implementation was organised as a small set of reusable modules and a collection of notebooks. Core configuration parameters, such as database connection details, cohort definitions and model hyperparameter grids, were centralised so that they could be adjusted without modifying multiple files.

5.2 Data ingestion and transformation outputs

The final data ingestion layer consisted of three Python scripts that interfaced with Apify:

- `ingest_apify_instagram.py`,
- `ingest_apify_tiktok.py`,
- `ingest_apify_facebook.py`.

Each script requested data from a corresponding Apify actor, downloaded the resulting JSON files, parsed the relevant fields and inserted them into the PostgreSQL `accounts` and `posts` tables. The scripts were designed to be idempotent. They updated existing account records when new follower information was available and added new posts that were not already present.

The primary outputs of this stage were:

- a populated `accounts` table containing one row per SME account with platform, country, sector and follower information,

- a populated `posts` table containing one row per post with timestamps, captions, content types and raw engagement metrics,

These transformed tables formed the stable input to all subsequent processing steps. No manual editing of the raw data was required after this point.

5.3 Feature and label generation outputs

Feature engineering and label construction were implemented in a dedicated Jupyter Notebook for each platform, backed by supporting Python modules that encapsulated reusable transformation functions. The notebooks executed the sequence described in Chapter 3 and wrote modelling ready tables for each cohort.

Using pandas and NumPy, the notebooks loaded posts and accounts from PostgreSQL, applied the account level and post level filters, computed engagement rates and constructed the log transformed and demeaned engagement variables. Textual, temporal, historical and account features were then added.

The main outputs produced by this stage were:

- one feature label table per cohort, stored as serialised files (for example parquet or pickle) and optionally as tables in a dedicated schema in PostgreSQL,
- summary statistics tables for each cohort, including post counts, basic descriptive statistics of engagement and class balance for the high engagement label,
- diagnostic plots that visualised the distribution of the raw engagement rate and the demeaned log engagement rate, which provided a check on the chosen transformations.

These artefacts provided a clear boundary between data preparation and modelling. The classification notebooks consumed only the feature label tables, not the raw scraped data.

5.4 Model training and evaluation outputs

The modelling layer was implemented in a set of Jupyter Notebooks, one per platform or group of related cohorts. Each notebook followed a common pattern:

1. load the feature label table for a given cohort,
2. sort posts chronologically and create temporal train and test splits,
3. define the Random Forest and XGBoost model instances with cohort specific hyperparameters,
4. fit both models on the training data with class weighting or positive class scaling,
5. generate predictions on the test set and compute evaluation metrics,
6. extract feature importance scores from each model.

Scikit learn provided the Random Forest implementation, train test split utilities and metric functions such as AUC and average precision (Pedregosa et al. 2011). XGBoost was used for gradient boosting with control over learning rate, tree depth and positive class weighting (Chen & Guestrin 2016).

The concrete outputs from this stage included:

- saved model objects for each cohort and algorithm, stored in a serialised format so that they could be reloaded for further analysis,
- metric tables containing AUC, average precision and top decile lift values for each cohort and model type,
- feature importance tables for Random Forest and XGBoost, listing the most influential predictors and their relative contributions.

These outputs fed directly into the evaluation chapter, where they were combined and compared across platforms, countries and sectors.

5.5 Visualisation and reporting outputs

The final implementation stage focused on visual and tabular reporting. In separate notebooks, the metric tables and feature importance outputs described above were merged and formatted for inclusion in the thesis.

Using Matplotlib, the following visual outputs were generated:

- ROC curves and precision recall curves for representative cohorts, showing the trade off between true positive rate, false positive rate, precision and recall,
- bar charts of top features by importance for selected models, highlighting which textual, temporal, historical and account variables most strongly influenced predictions,
- bar charts of top decile lift values across cohorts, which provided an at a glance comparison of where the models delivered the largest gains over random selection.

In addition, tables summarising performance by platform, by country and by sector were exported from the notebooks as LaTeX table code and integrated into the evaluation chapter. No graphical user interface or live dashboard was implemented, since the objective of the project was to develop and critically assess predictive models rather than to deploy a production system.

Overall, the implementation combined Python scripts, Jupyter Notebooks and a PostgreSQL backend into a coherent workflow that produced transformed data, trained models, evaluation metrics and visual summaries. These artefacts collectively form the empirical foundation for the analysis presented in the remainder of the thesis.

6 Evaluation

The purpose of this section is to provide a comprehensive analysis of the empirical results obtained from the classification models and to interpret the main findings in light of the research questions. The focus is on the extent to which the models can predict high engagement posts across platforms, countries and sectors, and on what the results imply for both academic understanding and SME practice. Only the most relevant results that support the research objectives are presented.

The analysis is organised as follows. Section 6.1 reports overall classification performance across sectors and platforms. Section 6.2 compares model performance by country and platform. Section 6.3 examines feature importance as an indicator of the drivers of engagement. Section 6.4 analyses timing and caption length effects using heatmaps and binned engagement plots. Section 6.5 synthesises these results and discusses their implications for research and practice.

6.1 Experiment 1: Overall classification performance by sector and platform

The first experiment evaluates the overall ability of the models to separate high engagement posts from other posts across sectors and platforms. Two complementary metrics are used. Figure 3 presents the test AUC scores, which measure ranking quality in a threshold independent way. Figure 4 shows the corresponding top decile lift values, which translate model performance into a more business oriented notion of gain.

Across all sectors and platforms, AUC values lie between roughly 0.65 and 0.85. This is well above the random baseline of 0.50 and indicates that the models consistently extract useful signal from the features. Instagram exhibits the most stable performance. For beauty, fashion and retail, Instagram AUC scores cluster around 0.81 to 0.83 with relatively narrow error bars. This suggests that, regardless of sector, the models can reliably rank Instagram posts by their probability of achieving above baseline engagement.

TikTok performance is more modest, with AUC scores in the 0.65 to 0.70 range for all three sectors. The error bars are somewhat wider than on Instagram, which points to greater variation across individual TikTok cohorts. Facebook shows a mixed picture. Beauty and fashion cohorts reach AUC values close to 0.70 but with larger uncertainty, reflecting the smaller and noisier Facebook sample. In contrast, retail on Facebook reaches the highest AUC of all platform sector combinations, around 0.84, indicating more regular engagement patterns in that particular context.

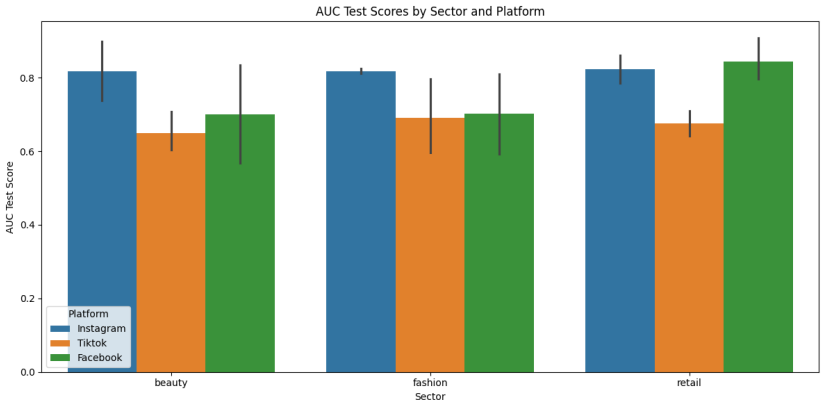


Figure 3: Test AUC scores by sector and platform for the best performing classifier in each cohort. Error bars indicate variability across the underlying country specific models.

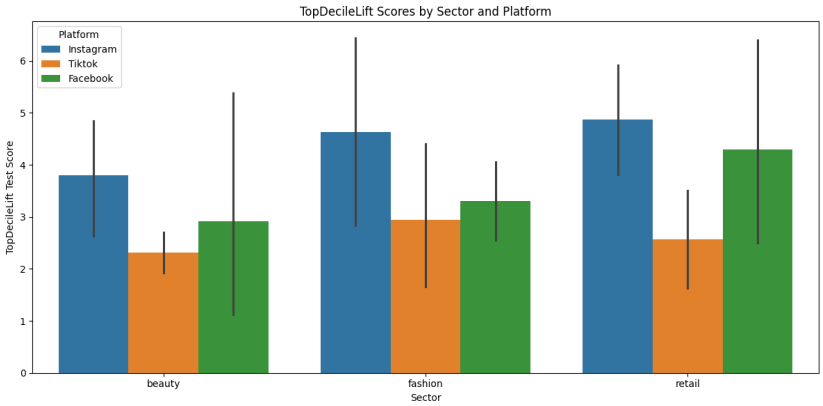


Figure 4: Top decile lift scores by sector and platform. Error bars indicate variability across the underlying country specific models.

The top decile lift results in Figure 4 confirm and extend these findings. For Instagram, lift values are around 3.8 for beauty, 4.6 for fashion and 4.9 for retail. In practical terms, this means that the ten per cent of posts ranked highest by the Instagram models contain nearly four to five times as many high engagement posts as would be expected under random selection. TikTok achieves lift values between about 2.3 and 3.0, which still represents a clear improvement over chance, although the gains are smaller and more variable than on Instagram.

Facebook again displays a mixed pattern. Beauty and fashion cohorts show moderate lifts of approximately 2.9 and 3.3 respectively. Retail on Facebook stands out, with a lift of about 4.3, which is in line with its strong AUC score. For SMEs in that sector and on that platform, a model guided selection of the top ten per cent of posts can therefore capture more than four times as many high performers as a naive strategy.

Taken together, these results show that the modelling framework delivers both statistically meaningful and practically useful performance. Instagram offers the most consistent gains across sectors, TikTok provides moderate improvements that may be enhanced with richer platform specific features and Facebook delivers high value in the retail segment but more uncertain results elsewhere. These overall patterns motivate the more detailed country level and feature level analyses that follow in the subsequent experiments.

6.2 Experiment 2: Comparison by country and platform

The second experiment examines whether predictive performance differs between Ireland and Nigeria once we fix the modelling approach. Figure 5 reports the test AUC scores for the XGBoost classifier

by country and platform, with error bars indicating the variation across sector specific cohorts. A corresponding plot for the Random Forest classifier shows very similar patterns and is provided for completeness in Figure 6.

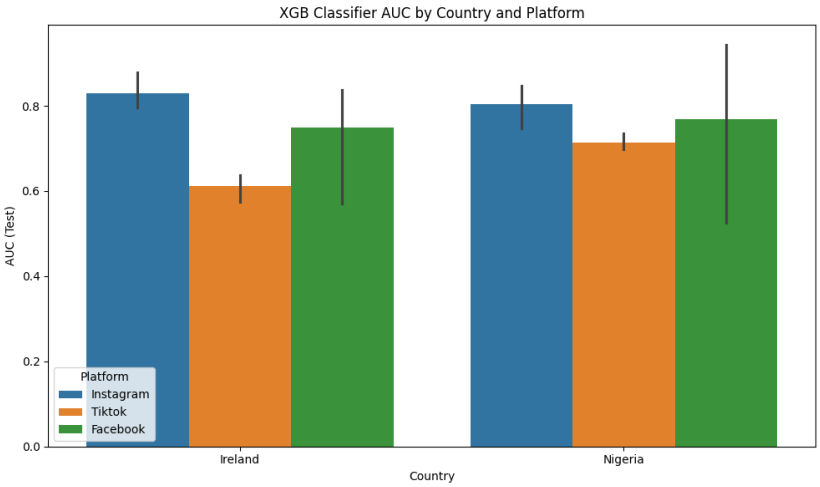


Figure 5: XGBoost classifier test AUC by country and platform.

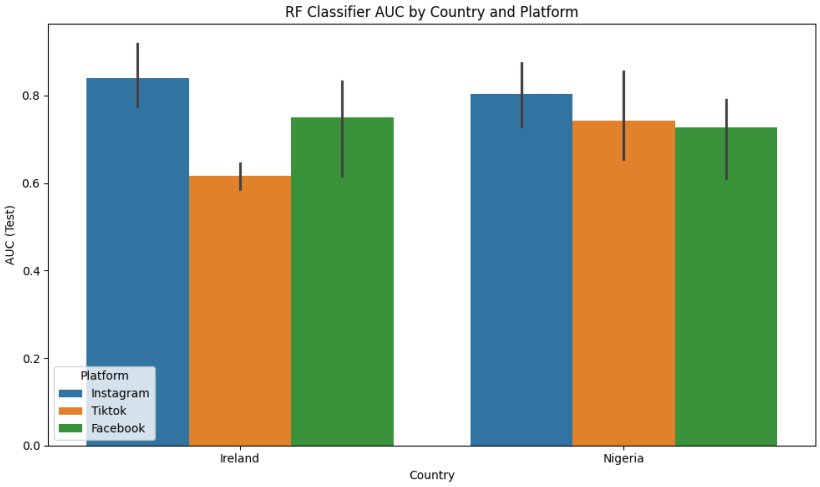


Figure 6: Random Forest classifier test AUC by country and platform.

Several observations emerge from Figure 5. First, Instagram is consistently the most predictable platform in both countries. In Ireland, the XGBoost AUC for Instagram is approximately 0.83 with a narrow confidence interval. In Nigeria, the Instagram AUC is only slightly lower at about 0.80. These values confirm that, regardless of national context, the feature set captures strong signals for ranking posts on Instagram and that the models can reliably discriminate between high engagement and other posts.

Second, TikTok shows a clear cross country contrast. In Ireland, TikTok AUC is around 0.61, only moderately better than random. In Nigeria, the TikTok AUC rises to roughly 0.71 and the error bars are relatively tight. This ten point increase suggests that engagement behaviour on TikTok is more structured and predictable for Nigerian SMEs than for Irish SMEs. One plausible interpretation is that TikTok is a more central platform for Nigerian SMEs in the sampled sectors, so content strategies and audience responses have converged on clearer patterns that the model can learn.

Third, Facebook displays mid range AUC values in both countries, around 0.75 in Ireland and 0.77 in Nigeria. The Nigerian estimate has a wide uncertainty band, which reflects the smaller volume of usable Facebook data and the noisier engagement distribution. Nonetheless, the point estimates

indicate that Facebook engagement is moderately predictable, particularly in sectors such as retail where Experiment 1 already showed strong performance.

Random Forest results in Figure 6 follow the same ordering. XGBoost tends to match or slightly exceed Random Forest AUC on each country platform combination, but the differences are small. This reinforces the conclusion that the main gains arise from good feature engineering and cohort specific labelling, with the choice between Random Forest and XGBoost primarily affecting fine tuning rather than the overall feasibility of the task.

From a comparative perspective, these findings provide quantitative evidence that platform effectiveness and predictability vary by country. Instagram emerges as a robust channel for both Irish and Nigerian SMEs. TikTok appears particularly promising for Nigerian SMEs, while its engagement patterns are less stable in Ireland. Facebook offers moderate predictability in both contexts, with especially strong results in Nigerian retail. These differences support the use of separate country platform models rather than a single pooled model and underscore the importance of tailoring social media strategies to local digital ecosystems.

6.3 Experiment 3: Feature importance and drivers of engagement

The third experiment investigates which features contribute most to the predictive performance of the models. For interpretability, feature importance scores from Random Forest (RF CLS) and XGBoost (XGB CLS) were normalised to lie between zero and one within each platform and averaged across the relevant cohorts. Figures 7, 8 and 9 display the top ranked features for TikTok, Facebook and Instagram respectively.

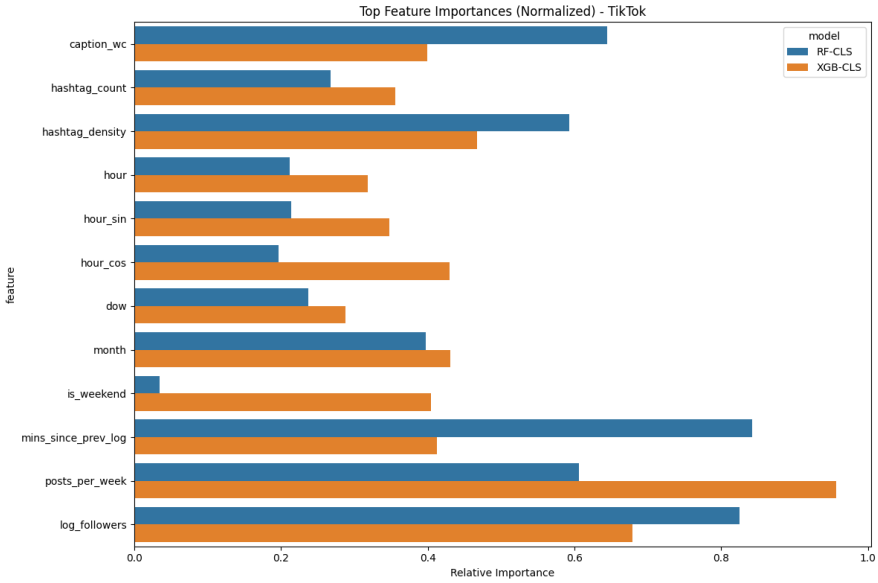


Figure 7: Normalised feature importances for TikTok, averaged across cohorts, for Random Forest and XGBoost classifiers.

On TikTok (Figure 7) the most important features are account scale and activity, together with caption structure and hashtag usage. Both models assign high importance to `log_followers`, `posts_per_week` and `mins_since_prev_log`, indicating that larger and more consistently active accounts are more likely to produce high engagement posts and that recency of posting matters. Caption word count and hashtag density are also prominent. The Random Forest in particular ranks caption length and hashtag density very highly, suggesting that richer captions combined with a focused set of hashtags help posts stand out in the TikTok feed. Temporal features (`hour`, `hour_sin`, `hour_cos`, `dow`) have moderate importance, which is consistent with Experiment 4 where timing effects are visible but not the only driver of engagement. Overall, TikTok engagement appears to be shaped by a combination of how often the SME shows up, the size of its audience and how much textual and hashtag information accompanies each video.

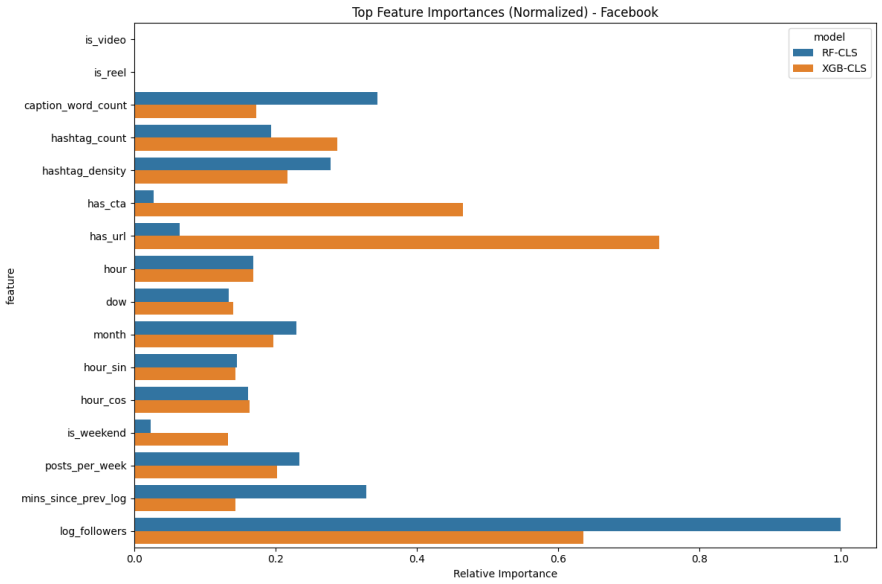


Figure 8: Normalised feature importances for Facebook, averaged across cohorts, for Random Forest and XGBoost classifiers.

On Facebook (Figure 8) the picture is slightly different. Here, `log_followers` dominates the Random Forest importance profile, reflecting the strong role of audience size in Facebook engagement. For XGBoost, two content level indicators, `has_url` and `has_cta`, emerge as very influential. Posts that include a clickable link or a clear call to action are much more likely to be classified as high engagement. This mirrors the broader marketing literature, which stresses that explicit prompts and easy paths to further information increase user response. Posting frequency (`posts_per_week`) and recency (`mins_since_prev_log`) also play a meaningful role, especially in the Random Forest model, while basic temporal features (`hour`, `dow`, `month`) contribute moderately. Content type flags (`is_video`, `is_reel`) appear in the importance rankings with smaller weights, indicating that video format helps but is not as decisive as follower base and explicit calls to act.

For Instagram (Figure 9) the feature importance profile combines elements of the TikTok and Facebook patterns. Account level variables `log_followers` and `posts_per_week` are consistently the top two features for both models, confirming that audience size and regular posting underpin strong performance. Recency (`mins_since_prev_log`) is also important, particularly for the Random Forest, which suggests that long gaps between posts tend to reduce the chance that a given post will be a top performer.

At the same time, several finer grained content and text features matter on Instagram. The presence of an external link (`has_url`) has substantial importance in the XGBoost model, while readability scores and caption word count carry meaningful weight in the Random Forest. Sentiment related variables (`sent_pos`, `sent_compound`) and subjectivity contribute modestly but non trivially, which points to the value of positive and engaging language. Hashtag density and hashtag count also appear among the top features, again favouring a balanced use of hashtags rather than either very sparse or very dense tagging. Emoji count, while less important than structural features, still receives non zero importance, hinting that expressive visual cues add marginal value.

Taken together, these feature importance results show that some drivers of engagement are common across platforms, while others are platform specific. For all three platforms, follower base, posting frequency and recency consistently rank near the top, which reinforces the idea that regular activity and long term audience building are foundational. TikTok places more emphasis on caption length and hashtag density, Instagram rewards a mix of structured captions, external links and subtle sentiment cues and Facebook is particularly sensitive to explicit calls to action and link presence. From a practitioner perspective, these findings identify concrete levers that SMEs can adjust when planning content on each platform. From an academic perspective, they provide quantitative support for theories that emphasise both reach building and message design as key components of social media marketing effectiveness.

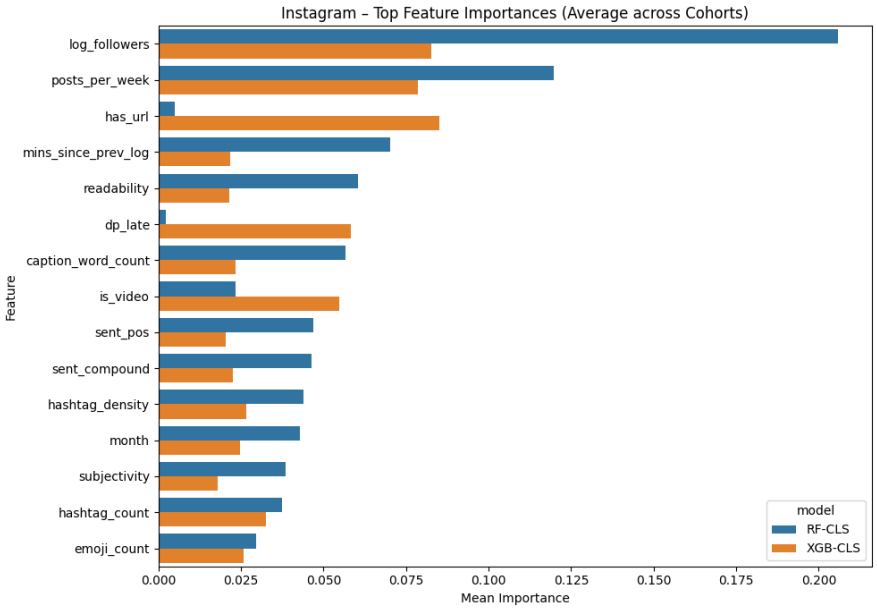


Figure 9: Normalised feature importances for Instagram, averaged across cohorts, for Random Forest and XGBoost classifiers.

6.4 Experiment 4: Timing and caption length effects

The fourth experiment uses descriptive visualisations to explore two controllable aspects of SME posting behaviour: when posts are published and how long captions are. These analyses do not introduce new models. Instead they summarise median engagement rate patterns, using the same engagement definition that underlies the classifiers. The aim is to understand whether the patterns that the models exploit are consistent with interpretable timing and caption effects.

6.4.1 Timing heatmaps

Figures 10 to 12 show timing heatmaps for median engagement rate by day of week and coarse time of day buckets for three representative cohorts: fashion SMEs on TikTok, retail SMEs on Facebook and beauty SMEs on Instagram, each split by country.

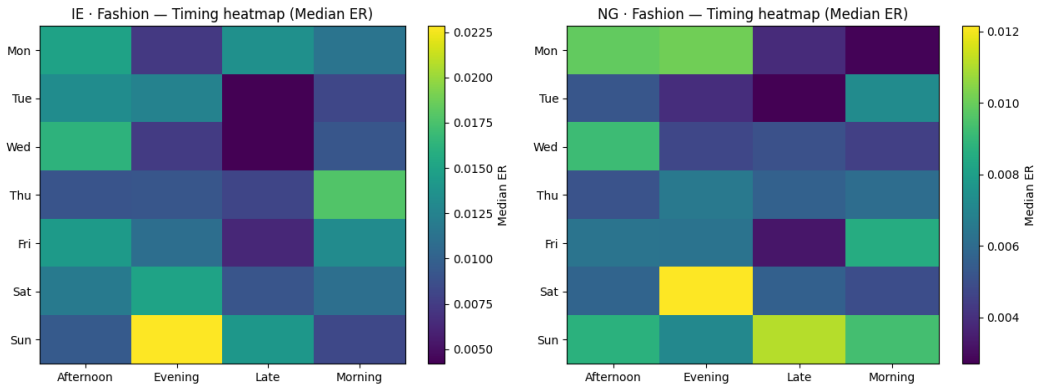


Figure 10: Timing heatmaps of median engagement rate for fashion SMEs on TikTok in Ireland (left) and Nigeria (right).

For fashion SMEs on TikTok (Figure 10) both countries show clear variation over the week and across time of day. In Ireland, engagement is elevated on Monday and Wednesday afternoons and on Sunday evenings, with a noticeable peak on Sunday evening. Thursday mornings also perform well,

while midweek late night slots tend to be weaker. In Nigeria, afternoons and evenings at the start of the week already yield reasonable engagement, but the strongest cells appear around late Friday and Saturday evenings, as well as late Sunday. This pattern suggests that Nigerian fashion audiences on TikTok are particularly active around weekend late slots, whereas Irish audiences respond strongly to Sunday evening content and selected weekday afternoon or morning posts.

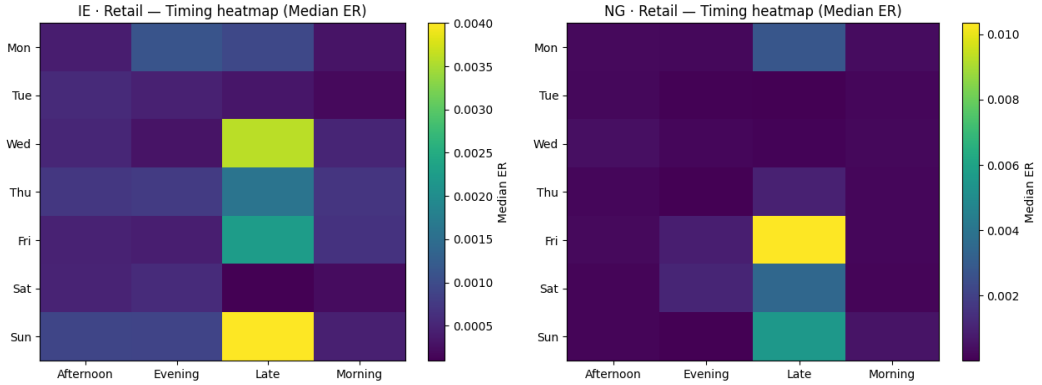


Figure 11: Timing heatmaps of median engagement rate for retail SMEs on Facebook in Ireland (left) and Nigeria (right).

Retail SMEs on Facebook (Figure 11) operate at much lower absolute engagement rates, but relative differences are still visible. In Ireland, most cells are uniformly low, with slightly higher engagement for late posts on Wednesday and Sunday. In Nigeria the pattern is more concentrated. There is a very clear spike for late Friday posts and a smaller, but still visible, elevation for late Sunday posts. Afternoon and morning slots remain relatively flat. This supports the view that, for Nigerian retail SMEs on Facebook, end of week late posts are the most promising windows for audience attention.

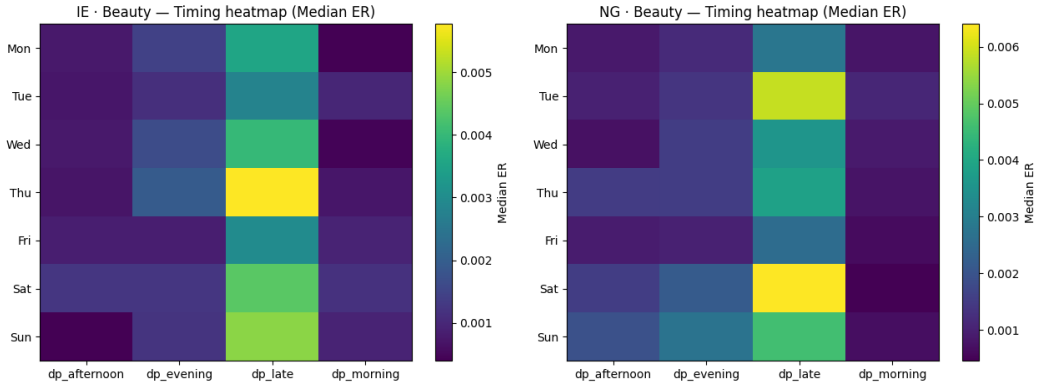


Figure 12: Timing heatmaps of median engagement rate for beauty SMEs on Instagram in Ireland (left) and Nigeria (right).

Beauty SMEs on Instagram (Figure 12) show the clearest structure. In Ireland, the late time of day bucket dominates, with the highest median engagement rates occurring for late posts on Thursday and Sunday and consistently above average values for late posts on most days. Evenings perform moderately well, whereas afternoons and especially mornings are comparatively weak. In Nigeria, the pattern is similar. Late posts again exhibit the strongest engagement, with peaks on Tuesday, Saturday and Sunday. Evening slots also perform reasonably, while mornings are almost always the least effective period.

Across all three platform sector combinations, late and evening posts tend to outperform afternoon and morning posts, but the best day of week combinations differ by country. These differences are consistent with the feature importance results, where hour and day of week variables carry non negligible importance. They also highlight that posting calendars should be tuned to local audience

routines rather than copied wholesale across markets.

6.4.2 Caption length and engagement

Figures 13 to 15 examine how median engagement rate varies with caption length. Captions are binned into word count intervals, and median engagement is computed within each bin for the same country platform sector pairs as above.

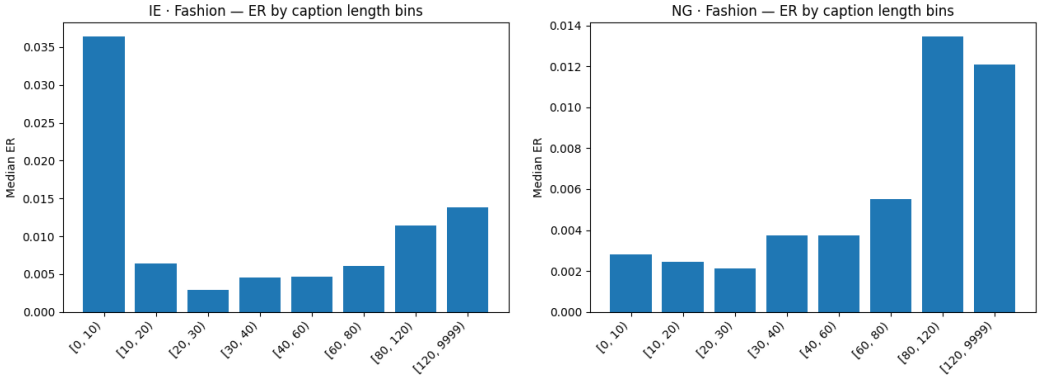


Figure 13: Median engagement rate by caption length bins for fashion SMEs on TikTok in Ireland (left) and Nigeria (right).

For fashion SMEs on TikTok (Figure 13) the relationship between caption length and engagement is non linear and differs by country. In Ireland, very short captions under ten words achieve the highest median engagement rate, although this bin likely contains relatively few posts and therefore should be interpreted with caution. Engagement then drops for captions in the 20 to 60 word range and rises again for long captions above roughly eighty words. In Nigeria, the pattern is more monotonic. Median engagement increases steadily as captions become longer, with the highest values for captions in the 80 to 120 word bin and only a slight drop for captions beyond 120 words. This suggests that succinct taglines can work in the Irish fashion context, but Nigerian fashion audiences on TikTok may respond better to more descriptive or narrative captions.

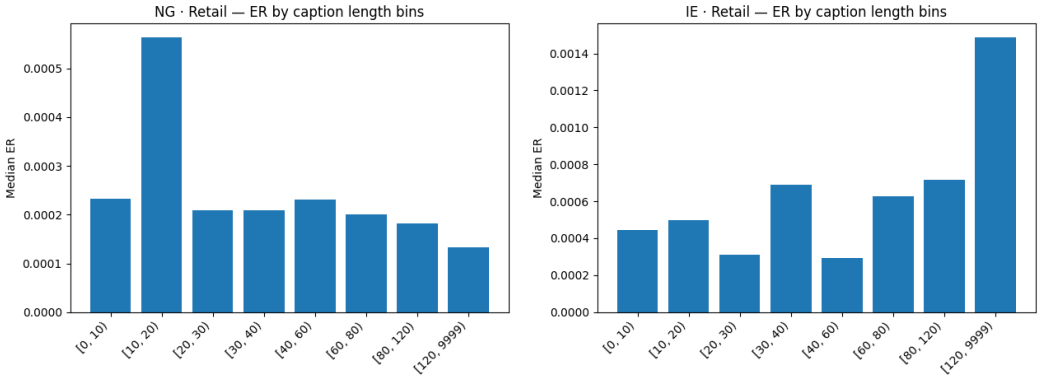


Figure 14: Median engagement rate by caption length bins for retail SMEs on Facebook in Nigeria (left) and Ireland (right).

Retail SMEs on Facebook show yet another pattern (Figure 14). In Nigeria, engagement peaks for captions in the 10 to 20 word range and is slightly lower for both shorter and longer captions. The drop for very long captions above 120 words suggests that long text may not be well suited to retail content on Facebook in this context. In Ireland, by contrast, the highest median engagement occurs in the longest caption bin above 120 words, with a secondary peak in the 30 to 40 word range. This implies that Irish retail audiences may be more willing to engage with detailed captions that explain offers or products, while Nigerian audiences favour shorter, more concise messages.

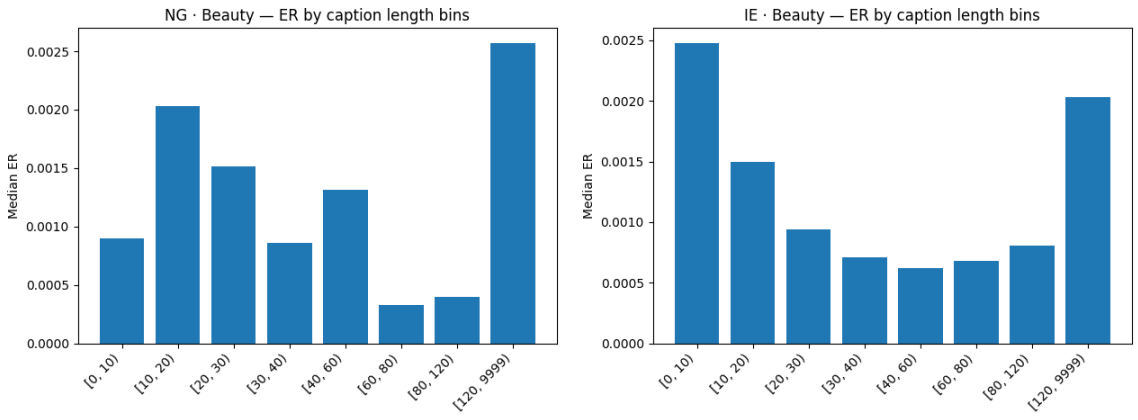


Figure 15: Median engagement rate by caption length bins for beauty SMEs on Instagram in Nigeria (left) and Ireland (right).

For beauty SMEs on Instagram (Figure 15) both countries exhibit a bimodal pattern. In Ireland, very short captions under ten words deliver the highest median engagement, followed by a decline for mid range captions between roughly 20 and 80 words and a noticeable increase again for very long captions above 120 words. In Nigeria, captions in the 10 to 20 word range and very long captions above 120 words both achieve relatively high engagement, while mid length captions perform less well. This pattern is consistent with a mix of styles on Instagram: either short, punchy taglines that complement strong visuals or longer, more informative captions that tell a story or provide detailed advice.

Overall, these caption length analyses confirm that the effect of text length on engagement is highly context specific and non linear. They also help explain why caption related features such as word count and readability appear among the important predictors in the feature importance experiment. The tree based models are able to capture these non linearities without imposing a simple increasing or decreasing relationship.

It is important to note that some bins, particularly those at the extremes of the length distribution, contain fewer posts and therefore have greater uncertainty around their median estimates. The patterns observed should therefore be interpreted as indicative rather than causal. Nevertheless, they provide practical guidance for SMEs. Irish and Nigerian firms can experiment with caption strategies that align with the shapes observed here, for example testing short and very long captions on Instagram rather than only mid length ones, or emphasising longer captions for Nigerian fashion content on TikTok.

6.5 Discussion

The experiments collectively show that the proposed framework can predict high engagement posts for SMEs with a level of accuracy that is both statistically meaningful and practically useful. The AUC and top decile lift results in Experiment 1 indicate that the models extract strong signal from the engineered features. Instagram is particularly consistent, with AUC values around 0.8 across beauty, fashion and retail and lift values close to four or five. For Facebook retail, the combination of high AUC and high lift suggests that engagement behaviour in that segment is especially regular. TikTok shows more modest but still valuable performance, especially in the Nigerian cohorts. These results support the view in the SME social media literature that platform choice matters and that some platform sector combinations offer richer and more predictable engagement opportunities than others (Fu et al. 2024, Matosas-López 2024, Prasetya et al. 2025).

Experiment 2 adds a cross country perspective. Instagram emerges as a robust channel in both Ireland and Nigeria, with only small differences in AUC between countries. In contrast, TikTok is considerably more predictable in Nigeria than in Ireland. This is consistent with qualitative accounts that describe TikTok as a rapidly growing and highly engaged platform for Nigerian consumers, and suggests that SMEs in that context have converged on clearer content patterns that the models can capture. Facebook shows broadly similar performance in both countries, but the wider uncertainty for the Nigerian estimate reflects the smaller sample of usable posts. Together, these findings reinforce the argument that digital strategies should be tailored to local contexts rather than imported wholesale,

a point also emphasised in cross country work on SME social media adoption (Palladan et al. 2023, Prasetya et al. 2025).

The feature importance analysis in Experiment 3 clarifies which variables drive these predictions. Across all platforms, follower base, posting frequency and recency consistently rank near the top. This aligns with prior evidence that long term audience building and regular posting underpin social media performance (Prasetya et al. 2025). At the same time, the models highlight platform specific drivers. On TikTok, caption length and hashtag density are particularly important, which suggests that SMEs benefit from captions that are either rich and descriptive or at least provide enough context to complement the video. On Facebook, explicit calls to action and the presence of links are strong predictors of high engagement, which is consistent with the platform's role as a traffic driver to external sites. On Instagram, a mixture of structural features such as caption length, readability and hashtag usage along with sentiment related features contributes to performance. These patterns provide quantitative support for theories that stress both audience reach and message design as determinants of social media effectiveness.

Experiment 4 shows that these model based findings are mirrored in observable timing and caption patterns. The timing heatmaps reveal that late and evening slots generally outperform morning and afternoon, but the best day time combinations differ by country and platform. For example, Nigerian fashion SMEs on TikTok achieve their highest median engagement with late posts on Friday and Saturday, whereas Irish fashion SMEs see strong performance on Sunday evening and selected weekday afternoons. Retail SMEs on Facebook in Nigeria benefit from very concentrated late Friday and Sunday posting, while beauty SMEs on Instagram in both countries show a clear advantage for late posts across multiple days. These results echo the importance of temporal variables in the models and add nuance by showing where in the calendar SMEs can focus their efforts.

Caption length analyses confirm that text design matters, but not in a simple linear way. For Nigerian fashion SMEs on TikTok, engagement rises steadily with caption length, with the highest rates for captions above eighty words. In Irish fashion, very short and very long captions perform best, with mid length captions less effective. Retail SMEs on Facebook in Nigeria see their best performance in the ten to twenty word range, while Irish retail benefits from the longest captions. Beauty SMEs on Instagram in both countries show a bimodal pattern, where very short taglines and very long narrative captions both work better than mid length text. These non linear patterns help explain why tree based models, which can flexibly partition the space, outperform simpler linear approaches (Breiman 2001, Chen & Guestrin 2016). They also provide concrete guidance for SMEs on the types of caption strategies that are worth experimenting with on each platform.

From an academic perspective, the study contributes to the literature in three main ways. First, it operationalises engagement in a way that controls for audience size and account baseline performance through the demeaned log engagement rate, then applies this consistently across platforms, sectors and two contrasting economies. This goes beyond much of the survey based SME work that relies on self reported performance metrics (Fu et al. 2024, Palladan et al. 2023). Second, it demonstrates that ensemble tree based models, when evaluated with temporally ordered splits and lift based metrics, can provide robust predictions of organic SME post performance, not just of paid advertisements. Third, it shows that platform and country specific patterns in timing and caption design can be captured quantitatively and linked back to interpretable features, which helps bridge the gap between technical modelling work and strategic marketing recommendations.

From a practitioner perspective, the findings offer actionable insights for SMEs. For Instagram, the models suggest that maintaining a steady posting rhythm, using either very concise or more narrative captions and timing posts for late slots on high engagement days can materially increase the chance that a post lands in the top engagement quartile. For Nigerian SMEs on TikTok, investing effort in richer captions and consistent weekend late posting appears particularly promising. For retail SMEs on Facebook, especially in Nigeria, aligning campaigns with late week late time windows and using concise captions with clear calls to action and links can yield significant gains. The top decile lift values translate these patterns into an intuitive metric: focusing promotion or boosting budget on the ten percent of posts that the models rank highest can deliver between two and five times more high engagement posts than spreading resources evenly.

At the same time, several limitations must be acknowledged. The analysis is restricted to three sectors and to SMEs that maintain public accounts on the selected platforms, which may introduce selection bias. Facebook data are relatively sparse, which increases uncertainty in that part of the

analysis. The models rely on structured features and simple text statistics and do not incorporate the visual content of posts, which is known to be important on image and video focused platforms. The study is observational rather than experimental, so it cannot definitively establish causal effects of timing or caption design. Finally, the data cover a specific time window, so platform algorithm changes or shifts in user behaviour could affect the stability of the findings.

These limitations point directly to avenues for future work. Incorporating image and video representations through computer vision models, expanding coverage to additional sectors and countries, and running controlled A/B tests that vary timing and caption strategies would help test the robustness and causal validity of the patterns observed here. Nonetheless, within these constraints, the experiments show that a relatively lightweight but carefully designed predictive framework can yield rich, transferable insight into how SMEs can use Instagram, TikTok and Facebook more effectively in different national contexts.

7 Conclusion and Future Work

This chapter concludes the thesis by revisiting the research questions, summarising the main findings and contributions, and outlining directions for future work.

7.1 Summary of aims and achievements

The overarching research question for this study was:

> How well can machine learning models predict high engagement social media posts for SMEs in Ireland and Nigeria, and how do the drivers of engagement differ across platforms and sectors?

To address this question, the thesis set four specific objectives.

1. To assemble a large scale, post level dataset of Instagram, TikTok and Facebook activity for fashion, beauty and retail SMEs in Ireland and Nigeria, and to define a robust engagement metric that controls for account size.
2. To engineer a set of textual, temporal, historical and account features that capture plausible drivers of engagement and to construct consistent high engagement labels across country platform sector cohorts.
3. To train and evaluate Random Forest and XGBoost classifiers for each cohort using temporally ordered validation, and to compare performance across platforms and countries.
4. To interpret the models through feature importance analysis and descriptive visualisations, and to derive actionable posting guidelines for SMEs.

The work has achieved these objectives to a substantial extent. A multi platform dataset of more than forty thousand posts was collected and stored in a reusable PostgreSQL schema. A demeaned log engagement rate was defined that removes account baseline differences while retaining post level variation. This provided a coherent basis for high engagement labels within each cohort.

Random Forest and XGBoost classifiers were trained on the engineered features and evaluated on held out, chronologically later data. Across cohorts, AUC values ranged from about 0.65 to 0.85 and top decile lift values from roughly two to five. Instagram showed consistently strong performance in both countries for all three sectors, and Facebook retail achieved particularly high AUC and lift scores. TikTok models were moderately accurate overall, with notably better performance in Nigerian cohorts than in Irish ones. These results confirm that the proposed methodology can predict high engagement posts with accuracy that is well above random and that is meaningful for content selection.

Feature importance analysis and timing and caption length visualisations provided insight into the underlying mechanisms. The models consistently identified follower base, posting frequency and recency as core drivers of high engagement, alongside platform specific factors such as caption length and hashtag density on TikTok, explicit calls to action and links on Facebook, and caption structure, readability and sentiment on Instagram. Timing heatmaps highlighted that late and evening slots on

particular days yield higher median engagement, with the best combinations differing by country and platform. Caption length effects were found to be non linear and context dependent, supporting the use of flexible tree based models.

7.2 Implications for research and practice

The findings have several implications for academic research. First, they demonstrate that a supervised learning framework built on behavioural data can complement survey based studies of SME social media adoption and performance. By working at the level of individual posts and controlling for account size and baseline performance, the thesis provides a granular view of engagement dynamics that is often missing from higher level studies.

Second, the results show that ensemble tree based methods, when combined with careful feature engineering and temporal validation, offer a robust approach to predicting social media outcomes in SME contexts. The cross country and cross platform comparisons illustrate how predictive performance and feature importance profiles vary with digital ecosystem characteristics, which adds nuance to theoretical claims about the universality of social media strategies.

For practitioners, especially SME owners and social media managers, the work offers concrete guidance. For example, the models suggest that Irish and Nigerian SMEs can improve their chances of achieving high engagement by:

- maintaining consistent posting schedules and avoiding long gaps between posts,
- paying attention to late and evening time slots on days where heatmaps show historically strong engagement,
- tailoring caption length to platform and context, using concise taglines or rich narrative captions on Instagram, longer captions for Nigerian fashion content on TikTok and shorter, focused captions with clear calls to action for Nigerian retail content on Facebook,
- incorporating explicit calls to action and links on Facebook where appropriate, and using hashtags in a targeted rather than excessive way.

The top decile lift values translate these recommendations into potential gains. Prioritising the ten per cent of posts that the models rate most highly can deliver between two and five times as many high engagement posts as a naive strategy that treats all posts as equal. In resource constrained SME environments, such improvements can have a meaningful impact on marketing efficiency.

7.3 Limitations

Several limitations should be acknowledged when interpreting these results. First, SMEs differ markedly in posting frequency, creative resources and existing brand strength. The models partly account for this through posting frequency and follower controls, but cannot fully separate content quality from underlying brand equity. Second, engagement outcomes are influenced by external events such as seasonal campaigns, economic shocks or offline promotions, none of which are directly observed in the dataset. Third, the Apify based collectors may miss deleted posts or posts that were only briefly visible, which means the data reflect a conservative snapshot of each account's activity rather than a complete history. Fourth, the analysis focuses only on three visually oriented sectors and may not generalise to business to business or service intensive SMEs. Fifth, although three platforms are included, other channels such as X or LinkedIn are outside the scope of this work and could be explored in future research. Sixth, the TikTok engagement weighting scheme and the percentile thresholds used for high engagement labels are based on domain informed choices rather than exhaustive sensitivity analysis. A more systematic exploration of alternative weightings and label definitions across cohorts would help to quantify how sensitive the conclusions are to these modelling decisions. Finally, the analysis covers a specific time period. Platform features and user behaviour evolve quickly, so the models may need regular updating to remain accurate.

7.4 Future work and potential for commercialisation

These limitations point to several avenues for meaningful future work. One direction is to extend the feature set with multimodal representations of visual and audio content using current computer vision and speech models. This would allow the framework to capture stylistic and aesthetic aspects of posts that are likely to influence engagement but are not visible in the present features.

A second direction is to link engagement predictions more directly to business outcomes such as website visits, enquiries or sales, where data are available. Combining social media analytics with transactional or web analytics data would enable models of revenue impact rather than engagement alone, building on studies that relate social metrics to financial performance.

Third, future research could move towards causal inference and experimentation. Controlled A/B tests that systematically vary posting time, caption style, hashtag usage or creative elements would help validate the patterns observed here and refine recommendations. Techniques from uplift modelling and contextual bandits could then be used to adapt posting strategies in real time.

A fourth avenue is to broaden the geographical and sectoral scope. Applying the framework to additional countries, platforms and SME industries would test its generalisability and might reveal further cross context differences. Comparative studies between developed and emerging markets could deepen understanding of how infrastructure and cultural factors shape social media effectiveness.

Finally, the framework developed in this thesis has potential for commercialisation as a decision support tool for SMEs. A practical system could integrate with platform APIs, automatically ingest recent posts, score planned content using updated models and suggest optimal posting times and caption strategies for each account. Such a tool could be offered as a lightweight subscription service or as part of a consultancy offering for SMEs that lack in house data analytics capacity.

In summary, this thesis has shown that predictive analytics can provide a structured and evidence based approach to SME social media strategy. While there is room for richer models and broader data, the results demonstrate that even relatively simple, interpretable features and tree based classifiers can deliver valuable insight into when and how SMEs should post on Instagram, TikTok and Facebook in Ireland and Nigeria.

References

Benslama, T. & Jallouli, R. (2022), 'Social media data analytics for marketing strategies: The path from data to value', *Journal of Telecommunications and the Digital Economy* **10**(2), 96–110.

Bogaert, M. & Delaere, L. (2023), 'Ensemble methods in customer churn prediction: A comparative analysis of the state of the art', *Mathematics* **11**(5), 1137.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Bruce, E., Shurong, Z., Egala, S. B., Amoah, J., Ying, D., Huang, R. & Lyu, T. (2023), 'Social media integration: An opportunity for SMEs sustainability', *Cogent Business and Management* **10**(1), 2173859.

Burez, J. & Van den Poel, D. (2009), 'Handling class imbalance in customer churn prediction', *Expert Systems with Applications* **36**(3), 4626–4636.

Casado Jiménez, E. (2025), 'Use of social networks in spanish hortofruit companies: Content analysis on instagram and tiktok', *VISUAL REVIEW. International Visual Culture Review / Revista Internacional de Cultura Visual* **17**(3), 197–216.

Chatterjee, S. & Kar, A. K. (2020), 'Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from india', *International Journal of Information Management* **53**, 102103.

Chen, T. & Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, New York, pp. 785–794.

Chen, W., Yang, K., Yu, Z., Shi, Y. & Chen, C. L. P. (2024), 'A survey on imbalanced learning: latest research, applications and future directions', *Artificial Intelligence Review* **57**(6), 137.

Durkin, M., McGowan, P. & McKeown, N. (2013), 'Exploring social media adoption in small to medium sized enterprises in ireland', *Journal of Small Business and Enterprise Development* **20**(4), 716–734.

Fu, C.-J., Silalahi, A. D. K., Yang, L.-W. & Eunike, I. J. (2024), 'Advancing sme performance: a novel application of the technological-organizational-environment framework in social media marketing adoption', *Cogent Business & Management* **11**(1), 2360509.

Haque, M. E., Shaik, S. & Borbely, Z. (2024), 'Leveraging social media for modern marketing: A comparative analysis of instagram and tiktok', *International Journal of Multidisciplinary Business Management* **12**(4), 33–50.

Harris, C. R., Millman, K. J., van der Walt, S. J. et al. (2020), 'Array programming with NumPy', *Nature* **585**(7825), 357–362.

Hunter, J. D. (2007), 'Matplotlib: A 2d graphics environment', *Computing in Science and Engineering* **9**(3), 90–95.

Imani, A. et al. (2025), 'Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning', *Engineering Applications of Artificial Intelligence* .

Johnson, K. & Kuhn, M. (2025), 'What they forgot to tell you about machine learning with an application to pharmaceutical manufacturing', *Pharmaceutical Statistics* **24**(1), e2366.

Kedi, W. E., Ejimuda, C., Idemudia, C. & Ijomah, T. I. (2024), 'Machine learning software for optimizing sme social media marketing campaigns', *Computer Science and IT Research Journal* **5**(7), 1634–1647.

Kemp, S. (2025*a*), 'Digital 2025: Ireland', DataReportal. Available at: https://datareportal.com/reports/digital-2025-ireland (Accessed: 2 December 2025).

Kemp, S. (2025*b*), 'Digital 2025: Nigeria', DataReportal. Available at: https://datareportal.com/reports/digital-2025-nigeria (Accessed: 2 December 2025).

Kluyver, T., Ragan Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S. & Willing, C. (2016), Jupyter notebooks: A publishing format for reproducible computational workflows, *in* 'Positioning and Power in Academic Publishing: Players, Agents and Agendas', IOS Press, pp. 87–90.

Kuhn, M. & Johnson, K. (2013), *Applied Predictive Modeling*, Springer, New York.

Lestari, R. B., Sulastri, Shihab, M. S. & Andriana, I. (2024), 'Social media marketing and its impact on SMEs' business performance', *KnE Social Sciences* **9**(14), 503–515.

Matosas-López, L. (2024), 'Can customer engagement in social media be used as a predictor of operating revenue in smes?', *Journal of Open Innovation: Technology, Market, and Complexity* **10**(4), 100380.

McKinney, W. (2010), Data structures for statistical computing in Python, *in* 'Proceedings of the 9th Python in Science Conference', Austin, Texas, pp. 56–61.

Palladan, A. A., AbdulKadir, K., Ahmed, I. S. & Abubakar, Y. (2023), 'Effects of social media adoption on sme performance: the moderating role of business strategy', *International Journal of Technoentrepreneurship* **4**(4), 297–320.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011), 'Scikit learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Plaatsman, S., Butt, M. & Dierikx, M. (2025), 'Machine learning based customer churn prediction in subscription publishing utilizing crisp dm methodology: An automated pipeline for multi publisher environments', *Preprints* . Preprint available at Preprints.org.

PostgreSQL Global Development Group (2023), *PostgreSQL Documentation*. Version 15. Available at: https://www.postgresql.org/docs/ (Accessed: 2 December 2025).

Prasetya, M. D., Ananda, P., Sembiring Ananda, B. K. F. & Fadli, F. (2025), 'The evolution of social media marketing: A comprehensive systematic review of current trends, challenges, and future insights', *Jurnal Pendidikan Teknologi Kejuruan* **8**(2), 103–126.

Rosário, A. T. & Dias, J. C. (2023), 'How has data-driven marketing evolved: Challenges and opportunities with emerging technologies', *International Journal of Information Management Data Insights* **3**(2), 100203.

Sanches, E. & Ramos, C. M. (2025), 'Evaluating the impact of instagram engagement metrics on corporate revenue growth: Introducing the loyalty rate', *Information* **16**(4), 287.

Santhiya, R. & Kannan, K. (2024), 'The impact of social media marketing and machine learning applications in digital advertising', *Journal of Philanthropy and Marketing* . Online first, May 2024.

Sharabati, A.-A. A., Ali, A. A. A., Allahham, M. I., Hussein, A. A., Alheet, A. F. & Mohammad, A. S. (2024), 'The impact of digital marketing on the performance of SMEs: An analytical study in light of modern digital transformations', *Sustainability* **16**(19), 8667.

Singh, M. (2024), 'Machine learning in marketing analytics', *International Journal of Enhanced Research in Management and Computer Applications* .