Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Driven Client Acquisition in Social Media: Optimizing Engagement through Convolutional and Recurrent Neural Networks

Andrey Sokolov

and.sokolov@innopolis.university

Universitet Innopolis https://orcid.org/0009-0005-3195-1655

Ekaterina Isaeva

Perm State University https://orcid.org/0000-0003-1048-7492

Ilya I. Kheyfetz

Perm State University https://orcid.org/0009-0006-4752-1526

Alexey Shutkin

Innopolis University https://orcid.org/0009-0007-7473-4508

Ivan A. Sychev

Perm National Research Polytechnic University https://orcid.org/0009-0009-6313-611X

Artur Sokolov

ITMO University https://orcid.org/0009-0008-9131-174X

Armen Beklaryan

National Research University Higher School of Economics https://orcid.org/0000-0001-5141-5140

Research Article

Keywords: client attraction, natural language processing; convolutional NN, recurrent NN, binary classification, data mining

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10588278/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

AI-Driven Client Acquisition in Social Media: Optimizing Engagement through Convolutional and Recurrent Neural Networks

Abstract: Despite significant interest in artificial intelligence, few studies have explored the use of AI tools specifically for client acquisition. However, since any commercial organization is interested in attracting new customers, the potential of AI can be employed for solving this task. This study investigates the application of neural networks and AI to handle diverse and complex datasets from VKontakte communities and food delivery services, leveraging large-scale data to enhance customer acquisition strategies. Experimental evaluation demonstrated that the proposed CNN architecture achieved 98.8% accuracy for target group classification and 90.6% accuracy for promotional post classification. LLM-based data augmentation improved the recognition of minority classes, while YaGPT generated promotional content consistent with the identified engagement patterns. Although the present implementation was developed for VK, the proposed architecture can be adapted to other social media platforms that provide comparable API access and textual user-generated content. Platform-specific modifications are required for data collection and preprocessing. The paper provides new insights into Internet marketing and the use of AI in it.

Keywords: client attraction; natural language processing; convolutional NN; recurrent NN; binary classification;data mining

1 Introduction

Neural networks (NN) are widely utilized in AI, especially in various online services, where they have proven highly effective. For instance, recommendation systems enhanced by neural collaborative filtering, effective mashup recommendations from big data in mobile networks, and mashup-oriented web API recommendations are areas that are being investigated [1], [2], [3], [4].

Recommendations enable the user to save time and make any service more convenient. However, one of the most important tasks for any service is attracting new customers. Despite the importance of this issue, to our best knowledge, previous studies have failed to address NN as the key attractors of new users.

Crucial tasks in attracting new clients are to find those users who will be potentially interested in a particular service and what data sources are effective for finding such users. With the development of technology, especially mobile devices, the number of such sources and services that aggregate social networks, application stores, instant messengers, etc. is also growing. However, crucial issues such as obtaining and preparing data from such sources for subsequent AI processing are not comprehensively described.

Despite the widespread adoption of artificial intelligence in marketing technologies, existing approaches are often limited to addressing customer retention challenges, like recommendation systems, or require the manual adjustment of targeting parameters.

To address the challenges of attracting new customers in today's competitive market, we have formulated the main research question:

RQ Main. “To what extent can convolutional and recurrent NN models improve automated client attraction on VKontakte (VK) by optimizing target audience identification, content generation, and placement?”

To answer this question, the following 5 research questions were formulated:

RQ1. “How accurately can NN models identify target groups in VK with limited training data?”

RQ2. “To what extent can NN models classify promotional post types in VK groups to support content targeting?”

RQ3. “To what extent can engagement metrics (likes, reposts, comments, and views) be used by neural network models to support subscriber growth prediction?”

RQ4. “To what extent can NN models improve the selection of VK groups for content placement and subsequent post tracking?”

RQ5. “How effectively can NN models generate advertising posts tailored to specific target audiences on VK?”

To address these research questions and develop an effective system to support customer acquisition, we propose a new NN architecture and methodology. Our work focuses on developing an automated NN model and supplementary algorithms to attract customers to VK communities and a food delivery application. This study

© The author(s) 2021. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

contributes to the application of AI in customer acquisition, demonstrating that AI-based approaches can enhance engagement across different business types. To achieve this, we designed a unique experimental architecture: a NN Model Incorporating External Artificial Intelligence Systems for Automatic Customer Acquisition. We conducted an in-depth analysis of models [21] [6] [15] [18], selecting key elements to inform our model's design. Leveraging CNN, RNN, and GRU architectures, our model optimizes potential client engagement and content targeting within VK communities, providing valuable insights into effective digital marketing strategies.

This work describes in detail the design of obtaining and preparing data and then presents a NN system for attracting new clients. The automated system contains its own NN architectures and built-in external NN, such as YaGPT [5]. Experimental results are presented when developing our own architectures.

The relevance of this research stems from the technical challenge of automating the full customer acquisition cycle: from collecting and tagging diverse, unstructured social media data to generating effective content. Existing approaches either lack sufficient flexibility to adapt to the dynamics of social media data, or require significant computational resources. The work addresses these issues by creating a hybrid architecture that combines the efficiency of convolutional and recurrent NN (CNN, LSTM, GRU) for feature extraction from short texts and the capabilities of large language models (LLM) for data balancing and post generation. This enables the automation of stages that previously required human expert involvement and ensures reproducibility through a careful selection of hyperparameters. The results can be embedded into the workflow for marketing automation and into data analysis in advertising.

This rest of the paper has been divided into the following parts: Section 2 reviews existing related works on data science and NN in marketing; Section 3 describes the main concepts and stages of our system, the methods used, and the system architecture; Section 4 presents experimental results at each step of the study, which are discussed in Section 5; the conclusions are drawn in Section 6.

Related Works

The literature review is divided into three parts: firstly, these are data sources and mechanisms for retrieving them: social network APIs, mobile application APIs, and their hybrids; the second group combines information about classification algorithms and types of NN in recommender systems; the third group combines information about related areas.

1.1 Data Sources and Extractions

Popular social networks are a good source of data because they contain comprehensive digital footprints of consumer preferences and purchasing patterns [6]. We use the VK API to collect data, which has not been utilised to date in the systems for automated client attraction.

Social networks have attracted considerable attention due to the need for efficient processing of a massive amount of generated data, embedded in various sources and at different scales [7]. These comprehensive and streaming data are not an easy target to collect, store, process, and extract knowledge. The task requires robust storage capacities, and efficient collecting and processing mechanisms.

While considering the architectures of existing systems for collecting and analyzing data from social networks [7], [8], [9], we come to the conclusion that the main challenge is that the social-network data are poorly structured. This raises a problem of the proper data processing when transferring them into the storage. To overcome this barrier, researchers use the ETL-process, which allows for retrieving information from the OLTP-systems or databases, its further conversion into the appropriate format, and uploading into the database. The data are collected from different social networks, where they can be presented in many formats and with various information loads. Therefore, the use of off-the-shelf software is not feasible. When developing this subsystem, other limitations of the data retrieved from social networks, namely, privacy issues, should be taken into account. The data should be collected for a certain period of time, which can show the dynamics of changes.

To build the models of user interests, public data from the social network, such as posts and groups features, are parsed or crawled, transferred to the model input, cleaned and processed, and stored in the required type in the database [10]. Web scraping techniques are also employed to obtain information from the Internet and convert it into structured content for further studies. When it is automated, it becomes less prone to errors, faster, and more efficient. Many intelligent bot programs have been deployed to capture massive amounts of data on the Internet, which makes this technique available to a wide range of people from a variety of industries [9].

1.2 Neural Networks and other Tools in Customer Acquisition

NN have attracted considerable attention in the context of social network studies due to the need to efficiently process the massive amounts of data generated by social media and to enhance machine learning techniques for improved learning from these data [16]. Most research in the field of social network data analysis focuses on text sentiment classification and content recommendations. The works [1, 2] demonstrate the effectiveness of NN in

recommendation systems, while studies [11, 12] propose methods based on knowledge graphs and long-term interests. However, these approaches do not address the problem of attracting new users or the processing of short text messages, imbalanced, or dynamic data typical of social networks.

Recent transformer-based language models have also been applied to marketing- and customer-related text analysis, BERT and RoBERTa were employed to cluster customer tweets for the creation of personalised landing pages. Although the study demonstrated the applicability of transformer architectures to customer segmentation, it did not evaluate their effectiveness for automated client acquisition. Similarly, paper investigated RuBERT and XLM-RoBERTa for Russian-language passenger review classification, reporting accuracies of 87.7% and 90.0%, respectively. However, the work focused on sentiment classification rather than customer acquisition or promotional content analysis. Although transformer-based models have demonstrated strong performance on a variety of NLP tasks, they generally require substantially larger computational resources and larger labelled datasets for effective fine-tuning. In the present study, the input data consist primarily of short group descriptions and promotional posts, for which lightweight convolutional and recurrent architectures provide an efficient trade-off between classification performance, computational cost, and deployment requirements. Therefore, the experimental comparison was limited to CNN, GRU, and LSTM models, which are better suited to the constraints of the proposed real-time client acquisition pipeline.

Several studies have shown that AI is a driver of purchasing power [6], [17], [18]. One of the tools for companies seeking to improve the quality of customer service is the social customer relationship management (Social CRM) [8], which comprises a set of processes that allow decision-makers to analyze customer data, in order to launch an efficient, customer-centric, and cost-effective marketing strategy. However, the inclusion of social media data in the CRM introduces new challenges, as it requires advanced analytical approaches to extract actionable insights from voluminous and heterogeneous data.

One of the main challenges in customer acquisition is how to extract, integrate, and prepare heterogeneous data for NN processing. Consequently, mashup technologies have become an important research direction, particularly with the development of mobile technologies and recommender systems [2].

A data mashup combines various data sources for effective analysis and processing. There are works where the accumulation of interests and the creation of a knowledge base are well thought out, for instance, the integration of streaming data into semantic mashups [19]. The use of API mashups, namely for traveller information support is considered in [1], web service recommendation through static mashup-API invocation relationships is attained in the motif-based graph attention for service recommendation (MGSR) [20]. Wu et al. [3] propose a neural framework based on multimodel fusion and multitask learning for mashup-oriented Web API recommendation, whereas [4] presents a probabilistic matrix factorization approach with implicit correlation regularization to solve the recommendation problem and enhance the recommendation diversity. In addition, AI-based approaches have been proposed for creating customer profiles during their physical presence in stores [18].

N. Ameen et al. analyse customer behaviour using the trust-commitment theory combined with AI in the absence of empirical studies on consumer experience [17]. Their findings are also relevant for the development of advertising content.

Nowadays, syndication of information channels represents a common form of mashup that combines content from multiple sources. Mashups are generally regarded as an integral component of the modern Web, where a variety of services are combined into a single application to provide the user with the best quality, convenience and efficiency.

Despite these advances, mashup-based approaches have not proved sufficiently efficient for processing rapidly changing data flows, and only a few studies address the generation of content that stimulates customer interest in a product. Furthermore, the existing literature lacks a comprehensive description of data processing pipelines, from collecting and filtering heterogeneous sources to class balancing under severe data imbalance. For this reason, the present study focuses on efficient convolutional and recurrent architectures that are well suited to short social media texts and practical deployment, while using LLMs specifically for data augmentation and content generation. Unlike classical approaches [16], which rely on static datasets and standard evaluation metrics, our research focuses on processing imbalanced social media data associated with the cold-start problem and proposes the use of LLM for sample augmentation.

2 Research Method

Since the purpose of this article is, firstly, to determine whether there is software capable of providing an appropriate data-stream, user, and source tag, which can be used as the input for a NN, and if not, to devise a method for obtaining it, and, secondly, to develop a NN that can attract a user to a group on a social network, we analyze

and partly use experimental designs of NNs described in articles [14], [21], [22], [23].

The proposed model follows an optimized three-stage client acquisition: (1) data extraction and preprocessing from VK groups via API and mashup tools, (2) identification of target groups through CNN and RNN models, and (3) predictive engagement analysis and automated post-generation using LLMs. Each stage is designed to maximize data integrity, enhance classification accuracy, and increase client acquisition effectiveness. The interaction of the modules is shown in Fig. 1.

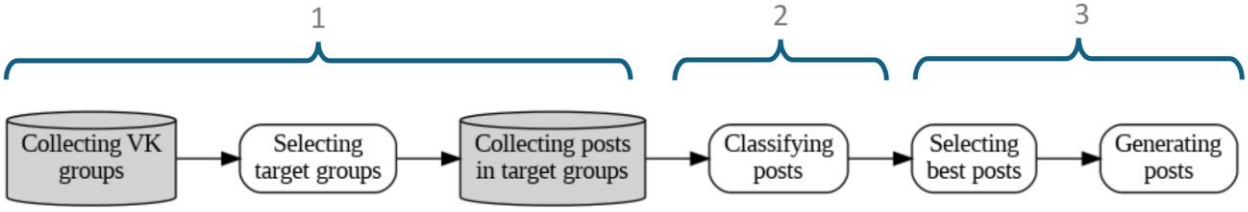


Fig. 1. Overall architecture of the proposed client acquisition pipeline: Module 1 – collecting data from the VK API and identifies target groups; Module 2 – classifying promotional posts and predicts engagement-related characteristics; Module 3 – generating new promotional content using an LLM

To obtain data for the customer acquisition model, social network APIs and mashups were used to retrieve historical data from social networks. To identify suitable software for social network data collection, the Targethunter [24] and Livedune [25] mashups were analysed. The data were collected with a custom Python application that sends HTTP requests to the REST API methods of the VK social network [26]. The primary advantage of this approach lies in its automation, which substantially reduces time required for data collection.

The main limitation of the system is that social network APIs do not support continuous data collection. The permitted number of requests per second and data availability vary across social networks, making initial data collection a relatively lengthy process. Fault tolerance is achieved by monitoring data collection errors. If an error prevents data retrieval, information about the issue is logged, and periodic attempts are made to obtain the required data block.

Finally, data integrity and confidentiality are crucial for the considered system. To ensure data privacy and compliance with platform policies, an anonymisation procedure is implemented during the data pre-processing stage, whereby users' personal identifiers (user_id) are removed and private information is stripped from message texts. VK API limits (primarily, limits on the number of requests per second) are handled via a batch processing mechanism and exponential backoff when limits are reached, which prevents access from being blocked. The proposed system operates exclusively on publicly available information obtained through the VK API in accordance with the platform's access policies. All personal identifiers are removed during preprocessing, and no private user information is processed. The generated content is intended to assist marketing specialists in content creation rather than replace human decision-making.

2.1 Collecting VK Groups

The search for potential target groups proceeded as follows: an initial list of delivery groups was compiled, and users were extracted from these groups. For each selected user, we retrieved lists of subscribed groups. This process was automated with a Python 3 script using libraries: *VK, time, datetime, requests, pandas, and sqlalchemy*.

2.2 Selecting Target Groups

After that, food delivery groups were selected from the list of received groups. The connection with food delivery was determined with a NN model based on the description of the group. To train the model, the data were prepared and annotated. The text descriptions of the groups were tokenized using the Keras tokenizer and reduced to one length. The annotation was performed manually by domain experts according to predefined labeling guidelines. Disagreements were resolved through discussion. To select the most accurate model, an experiment was conducted using a CNN, a recurrent long short-term memory (LSTM) network, and a gated recurrent unit (GRU) network.

The CNN architecture comprises an embedding layer, a one-dimensional convolution (Conv1D), a global pooling layer, and fully connected layers. Hyperparameter tuning was performed using the *Keras Tuner library*. For Module 1, the best results were achieved by models with an embedding size of 16, 64 filters for kernels of sizes 3, 7, and 9, and the ELU activation function. The Nadam optimiser with a learning rate of 1e-3 demonstrated the best convergence. The use of Dropout (0.5) helped prevent overfitting.

The GRU model was used to account for sequence context. The optimal configuration included 64 recurrent blocks, the ELU activation function and the Nadam optimiser. Dropout (0.4) was applied to prevent overfitting.

Like GRU, the LSTM model is also designed to process sequences. The best configuration comprises an

embedding size of 32, 64 LSTM blocks, and the SELU activation function. Additionally, for group classification, several LLMs, namely Mistral, Llama, and Qwen, were tested following fine-tuning. However, they were outperformed by specialised CNNs on the given dataset.

2.3 Collecting Posts in Target Groups

At the next stage, posts are collected from the selected target groups and the growth of the group's subscribers is monitored. To do this, a Python 3 script was implemented to interact with the VK API.

In addition to the above method, data import from mashup was also applied. Using the web interface of the Livedune system, two data sets were obtained: the set of posts in target groups and dynamics of the number of subscribers. Both datasets were merged into one with Python 3 script.

Both methods resulted in a dataset containing all properties necessary for post analysis. This dataset was saved to a pandas.DataFrame object for further operations.

2.4 Classifying Posts

We use a NN model to classify the generated posts. Also, to determine the optimal architecture, we have built a CNN, a recurrent LSTM network, and a recurrent GRU network. The data were manually annotated, and the texts of the posts were tokenized prior to processing. Additionally, several LLMs, including *Mistral-7B-v0.3*, *Qwen2-7b-instruct*, *LLama-8b-Instruct*, and *Suzume-llama-3-8b-multilingual-orpo-borda-half*, both in their original and fine-tuned versions, were used to further categorize the posts, .

2.5 Selecting the Best Posts

The quality of the post was assessed based on the increase in the number of subscribers, likes, reposts, comments, and views. Then the posts were marked up manually. The text of the posts was tokenized. The best solution was determined with the custom CNN, recurrent LSTM, and recurrent GRU network architectures. The same structure was used as in the models above to search for target groups. The NN hyperparameters were optimized using *Keras Tuner*.

2.6 Generating Posts

YandexGPT and other LLMs were further trained using the *Fine-tuning* method to generate advertising posts. For further training, the best posts were selected based on the assessment that was conducted in the previous module. The data was prepared in the JSON lines format.

The complete source code, preprocessing scripts, datasets, and model configurations are publicly available in the GitHub repository to facilitate reproducibility of the reported experiments.

3 Experimental Results

4.1 Experimental Setup and Data

The Experiments we conducted separately for each of the three modules: Module 1 (group classification), Module 2 (post classification), and Module 3 (post generation). Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC. For the classification tasks, the dataset was split into training, validation, and test sets. The train, validation, and test partitions remained fixed throughout all experiments to ensure a fair comparison between the evaluated models.

For group classification (Module 1), we dealt with the binary classification task. The initial corpus consisted of 11,613 records. After removing empty entries and cleaning the data, the sample was reduced to 8,500 instances: 5,000 target groups (food delivery), 3,500 non-target groups (other interests and businesses). To improve classification quality, the non-target class was expanded with 1,500 difficult negative examples (recipes, aggregators, non-food deliveries). The final balanced dataset comprised 10,000 vectorised representations.

For post classification, we formulated a five-class classification problem. Class 0 corresponds to soft-sell posts, i.e., promotional posts intended to increase brand awareness and customer engagement without offering explicit customer incentives. Classes 1–4 represent posts offering discounts, gifts, cashback, and on-time delivery, respectively. Due to a severe class imbalance (for example, the ‘Cashback’ class accounted for <3% of the original dataset), data augmentation was applied using an LLM (Gemma-4-31b-it). The underrepresented classes were segmented into 10 clusters each, and posts were generated in equal proportions for every cluster to balance the sample to 1,000 examples per class. All LLM-generated samples were included exclusively in the training set,

whereas the validation and test sets contained only manually annotated original posts. The final set consisted of a training sample of 3,500 posts, a test sample of 500, and a validation sample of 1,000. The datasets are available in the GitHub repository (see Section 4.5).

4.2 Module 1 Results: Target Group Classification

A comparative analysis of the models (CNN, GRU, and LSTM) demonstrated the highest performance of the CNN model with accuracy = 98.8%, precision = 98.81%, recall = 98.8%, F1-score = 98.8%, and AUC = 99.93%. These were achieved with the parameters: an embedding layer with 16-dimensional vectors, three parallel Conv1D branches with kernel sizes of 3, 7, 9 with 64 filters, ELU activation, a Dropout rate of 0.5, and the Nadam optimiser with a learning rate of 1×10^{-3}.

The GRU model showed a lower performance reaching accuracy, precision, recall, and F1-score = 98.0%, and AUC = 99.58%. While the LSTM model demonstrated the worst accuracy, precision, recall, and F1-score = 97.9%, with AUC = 99.73%, which outperformed the respective LSTM metric (Fig. 2).

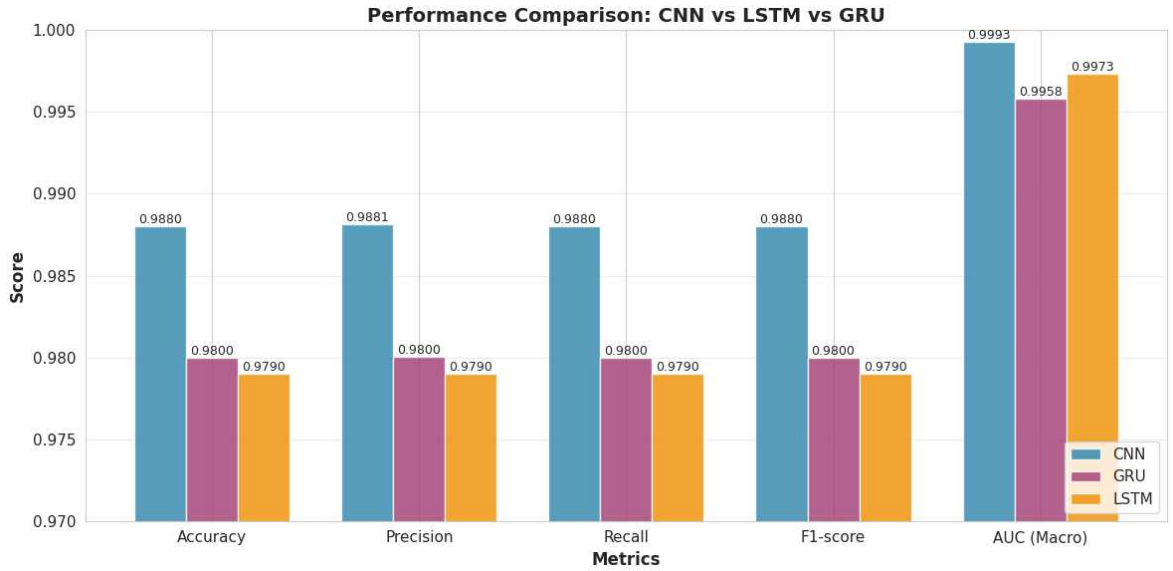


Fig. 2. Comparative visualisation of model metrics when solving group classification problems

The results show the CNN model's ability to extract local features from short group descriptions. The use of the dropout (0.5) and early stopping helped to avoid overfitting and eliminate the previously observed anomaly where test accuracy exceeded validation accuracy. The results are now stable, with the test set results yielding performance comparable to that on the validation set, confirming the model's generalisation ability.

These findings are further supported by the confusion matrix and ROC analysis shown in Fig. 3.

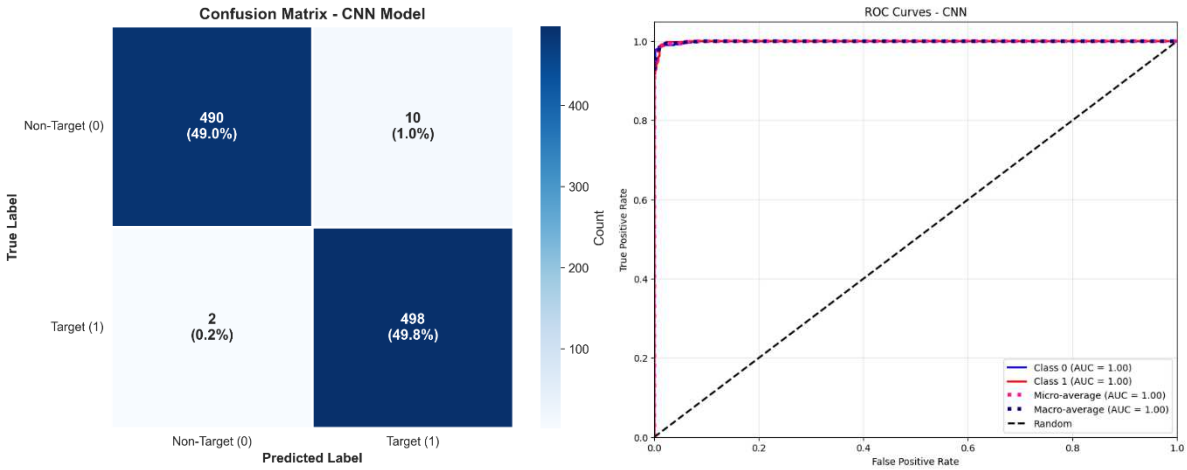


Fig. 3. Confusion matrix and ROC curves for the best-performing CNN model in solving group classification

tasks

Fig. 3 presents the confusion matrix and ROC curves for the best-performing CNN model. The confusion matrix shows that the model correctly classified 490 of 500 non-target groups and 498 of 500 target groups. Only 12 misclassifications were observed, including 10 false positives and 2 false negatives, resulting in an overall accuracy of 98.8%. A low number of false negatives indicates that the model rarely fails to identify target groups, which is particularly important for the subsequent customer acquisition pipeline.

The ROC curves are concentrated near the upper-left corner of the plot, indicating excellent separability between the two classes. The AUC values further confirm the robustness of the proposed classifier.

4.3 Module 2 Results: Post Classification

The multi-class post classification turned out to be more challenging due to the semantic proximity of the classes. As shown in Fig. 4, the CNN model achieved the best overall performance, reaching accuracy = 90.6%, precision = 90.86%, recall = 90.6%, F1-score = 90.68%, and AUC = 98.59% (Fig. 4).

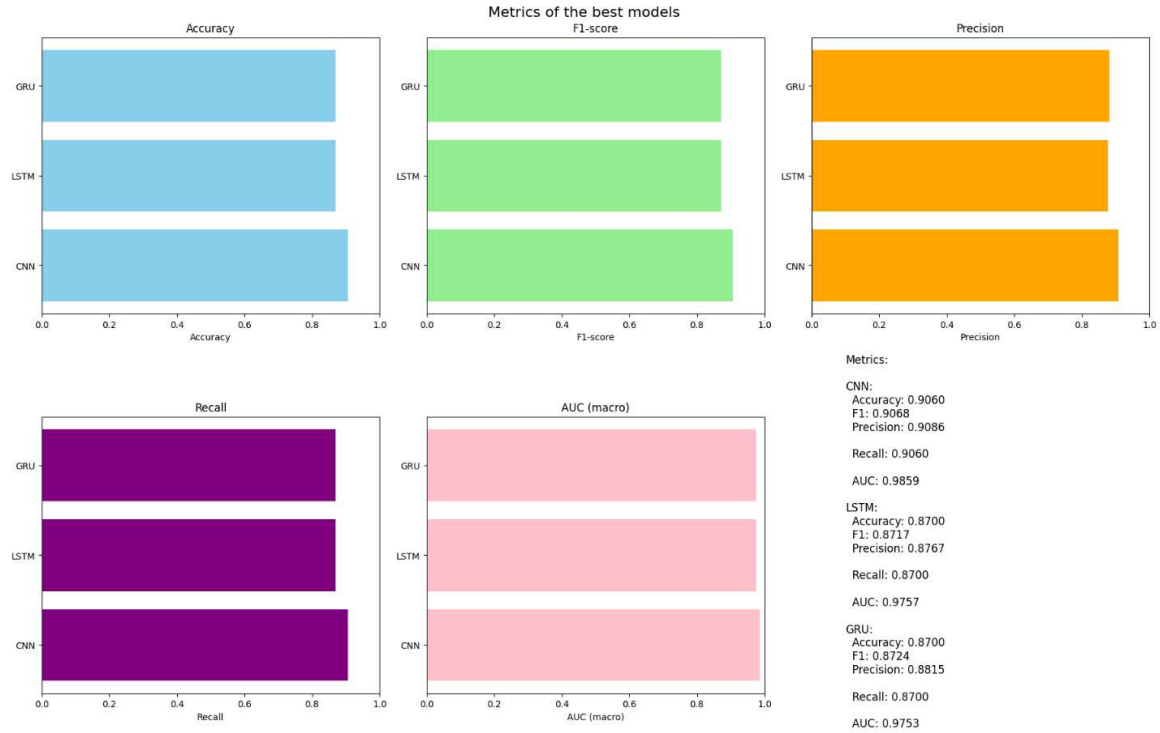


Fig.4. Comparative visualisation of model metrics in solving post classification tasks

The best-performing configuration comprised an embedding layer with 48-dimensional vectors, three parallel Conv1D branches with kernel sizes of 3, 4, and 5, ReLU activation, a Dropout rate of 0.5, and the Nadam optimiser with a learning rate of 1×10^{-2}. By comparison, the GRU and LSTM models achieved similar performance, both reaching accuracy = 87.0%, with F1-scores = 87.24% and 87.17%, respectively. Their AUC values (97.53% for GRU and 97.57% for LSTM) indicate that the recurrent architectures also learned meaningful representations, although they consistently underperformed the CNN.

The confusion matrix and ROC curves for the best-performing CNN model are presented in Fig. 5.

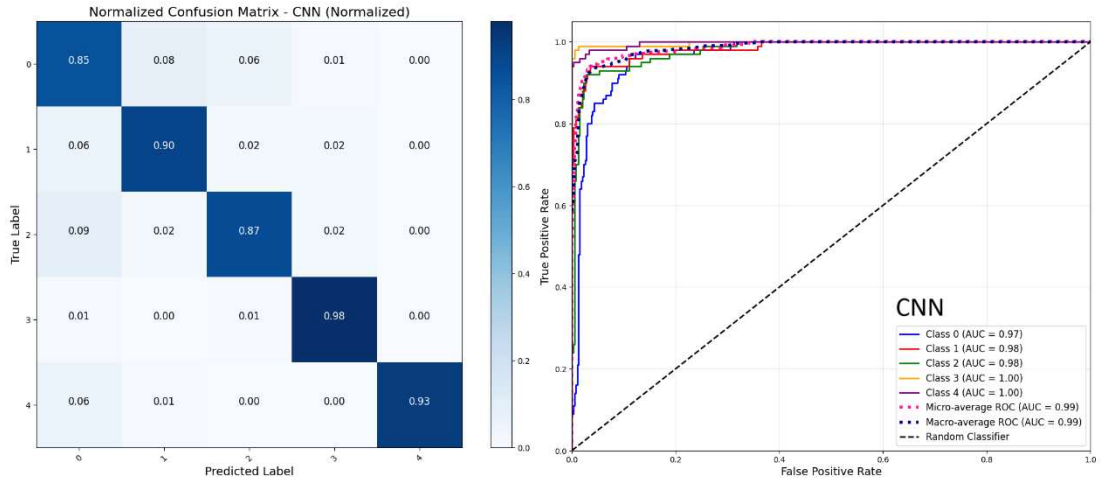


Fig. 5. Confusion matrix and ROC curves for the best-performing CNN model in solving post classification tasks

The confusion matrix indicates that the highest proportion of misclassifications occurred for the soft-sell class (Class 0). Approximately 8% of soft-sell posts were classified as discount posts and 6% as gift posts, whereas confusion with the cashback and on-time delivery classes was negligible. Similar confusion patterns were observed in the opposite direction, reflecting the semantic similarity between soft-sell content and promotional posts offering discounts or gifts. In contrast, the cashback and on-time delivery classes were recognised with high accuracy after LLM-based augmentation, despite being severely underrepresented in the original dataset.

The ROC curves further confirm the robustness of the classifier. The AUC exceeded 0.97 for every class, while the micro- and macro-averaged AUC reached 0.99, indicating excellent separability across all five classes. Together with the improved recognition of the originally underrepresented cashback and on-time delivery classes, these results suggest that LLM-based augmentation using Gemma-4-31b-it effectively mitigated the effects of severe class imbalance while preserving high classification performance for the remaining categories.

The superior performance of CNN across both modules is likely explained by the short length of VK group descriptions and promotional posts, where local lexical patterns play a more important role than long-range dependencies.

4.4 Module 3 Results: Post Generation

The third module was designed to generate advertising posts based on the promotional patterns identified by the classification models. Unlike the previous modules, the objective was not to classify existing content but to automatically produce new posts corresponding to a specified promotional strategy.

The prompts supplied to the LLM contained the target promotional strategy (soft-sell, discount, gift, cashback, or on-time delivery), while the generated posts followed the stylistic patterns extracted from the training corpus. Four LLMs were evaluated for post generation: YaGPT, Qwen2-7B, Llama3-8B, and Suzume-Llama-3-8B.

YaGPT generated the most consistent promotional texts, correctly incorporating explicit customer incentives and calls to action. Llama3-8B also produced coherent advertisements but occasionally introduced additional contextual information not specified in the prompt. Qwen2-7B demonstrated lower prompt adherence, sometimes generating content unrelated to the requested promotional scenario. Suzume-Llama-3-8B successfully followed the requested structure but occasionally produced incomplete or noisy outputs. Representative examples are shown in Table 9.

Table 9. Examples of prompts with target promotional strategy and responses of pre-trained LLM

Llm model	Prompts	Response
YaGPT	"Generate an advertising post for a food delivery group that promises the user a discount."	"Hi everyone! We have great news for all our customers! 😍 We're offering a special discount on all your orders! 🔥 Use the promo code [your promo code] and get 10% off your order! 🤩 This offer is valid until [date of expiration]. Hurry up! Your discount is waiting for you! 😉 FoodDelivery Discount SpecialOffer"
Qwen2-7b	"Generate an advertising post for a food delivery	"It's time to celebrate #InternationalWifeDay! Thank you for being the most beautiful part of our lives❤️ To all the wives out there, we wish you only the

	group that promises nothing to the user."	best! We are grateful for your love, care and support. You are the best! Don't forget to send your love to your wife today! 🤩 "
Llama3-8b	"Generate an advertising post for a food delivery group that promises the user a gift."	"Is your birthday coming up? Or maybe it's already today? 😻 Celebrate this wonderful holiday in the circle of relatives and loved ones! And about the preparation of a delicious feast will take care of us❤ We give all birthday boys Mini Set! With a minimum order of 1200 rubles🎁! 🍱What is included in the set? - Philadelphia Light with shrimp - Fried roll with chicken Promo code: DR **The promotion is valid 3 days before and 3 days after the event."
suzume-llama-3-8b	"Generate an advertising post for a food delivery group that promises the user a deliver on time."	"Important!!! Our express order for delivery of Chicken Cookie ER with Chicken is now being given to our birthday kids 🎉 To get a promo code to oblige us to deliver your order on time! ⚡Order through the app or to our website: [https://vk.cc/cn7BHM] (https://vk.cc/cn7BHM)"

As illustrated in Table 9, the generated texts preserved the requested promotional strategy and included typical marketing elements such as calls to action, promotional vocabulary, and emotional markers (e.g., emojis). The outputs required no manual template construction, demonstrating the feasibility of using LLMs for automated content generation.

Although all evaluated LLMs were capable of generating promotional content, YaGPT demonstrated the highest prompt adherence and the lowest frequency of irrelevant responses. This behaviour is likely related to its optimisation for Russian-language content, whereas the multilingual models occasionally produced semantically inconsistent outputs.

4.5 Addressing Research Questions (RQ Mapping)

The experimental results provide answers to the research questions formulated in the Introduction.

- RQ1 (Target Groups): The CNN model identified target groups with an accuracy of 98.8%, demonstrating that automated audience targeting on VK is feasible even when training data are limited.
- RQ2 (High-Engagement Posts): The CNN model achieved an F1-score of 90.68% in distinguishing between promotional post types, demonstrating the feasibility of automatically identifying content patterns associated with customer engagement.
- RQ3 (Subscriber Growth): The proposed pipeline supports subscriber growth prediction by combining engagement indicators with post classification. The high recall of the CNN model reduces the likelihood of overlooking potentially successful promotional content.
- RQ4 (Placement): The data collection module, together with the group classification model, automates the identification of suitable VK communities for content placement and subsequent monitoring.
- RQ5 (Generation): Module 3, based on YandexGPT, generates advertising posts according to the promotional patterns identified in the previous modules (examples are provided in Table 9), demonstrating the feasibility of automated promotional content generation.

The complete implementation of the proposed pipeline, together with the datasets and source code, is available in the Github repository (https://github.com/AndreySokolovInterprogram/AI-Driven-Client-Acquisition-in-Social-Media).

4 Discussion

An important observation is that the posts generated by the system closely match the characteristics of high-engagement posts, indicating that the proposed pipeline successfully transfers the patterns identified during the classification stage to the content generation module and aligns with client attraction strategies.

Unlike prior studies that focus solely on data extraction or analysis, this work provides a comprehensive data mining and client acquisition algorithm, integrating both data processing and NN deployment for VK. This contribution is particularly important because reliable data acquisition forms the foundation of the entire analytical pipeline. There are works, for instance [9], that describe data extraction and processing, in particular the tokenization of the textual data and the web scraping algorithm. However, their algorithms are not described in detail and are hard to replicate.

At present, the proposed system has been implemented only for VK. Although the semantic meaning of user interactions is largely consistent across social media platforms, API structures, field names, authentication mechanisms, and data collection procedures differ substantially. Consequently, dedicated data acquisition modules

require customisation for each platform. Furthermore, the VK API imposes request rate limits and privacy restrictions that constrain continuous large-scale data collection. Addressing these limitations may require distributed data storage, asynchronous batch processing, and more sophisticated data management strategies.

The previously observed discrepancy between validation and test accuracy was eliminated by introducing hard negative examples into the training dataset together with dropout (0.5) and early stopping. These modifications improved the stability of the training process and reduced overfitting. The reported results are based on a single train/validation/test split. Future work will include repeated experiments with cross-validation and statistical significance testing to further evaluate the robustness of the proposed approach.

The practical aspects of deployment are as follows. When integrating the system in real time, VK API limits must be taken into account. Accordingly, an architecture utilising queues (batch processing) and dynamic token renewal is proposed. Data confidentiality is ensured at the pre-processing stage by removing users' personal identifiers and hashing key fields before feeding the data into the NN.

A limitation of the post generation module should also be taken into account – several LLMs occasionally generated irrelevant contextual information or promotional details that were not explicitly requested in the prompt. Therefore, human verification remains advisable before such content is published or used in real-world marketing campaigns.

Future work will evaluate the real-world effectiveness of the generated posts in customer acquisition campaigns and extend the proposed pipeline to additional social media platforms.

5 Conclusion and Future Work

This study demonstrates the feasibility of automating the main stages of client acquisition. Expert involvement is required primarily during the initial definition of target groups, after which the remaining stages of audience identification and content processing can be performed automatically.

The proposed system has been implemented for client acquisition in VK food delivery communities and is currently being piloted within the workflow of the Inter Program company.

This study presents a new approach to facilitating customer acquisition using AI, and demonstrates that the CNN architecture reaching the accuracy of 98.8% for groups classification and 90.6% for post classification consistently outperforms the recurrent methods. The use of an LLM (Gemma-4-31b-it) for data augmentation successfully resolved the issue of severe class imbalance in the post classification task. The system has been piloted by the Interprogramma company, confirming the feasibility of automating the stages of targeting, content analysis, and advertisement generation, thereby reducing marketing costs and improving the effectiveness of advertising campaigns.

Although the current implementation is limited to VK, the proposed architecture can be adapted to other social media platforms that provide appropriate API access, subject to platform-specific modifications.

The experiments were conducted using a fixed train/validation/test split rather than a time-based partition. Future work will evaluate temporal generalisation using chronological data splits. Future work will also focus on extending the system to additional social media platforms, improving data acquisition from heterogeneous sources, and evaluating the effectiveness of generated advertising content in real customer acquisition campaigns. Additional modules for automatically selecting suitable publication communities will also be incorporated.

Acknowledgment

This research was supported by the LLS "Interprogram". We also express our gratitude to the Faculty of Mechanics and Mathematics of Perm State University for providing the computing power of the "Kepler" multiprocessor computing complex.

References

[1] B. Karakostas and Z. Kalamboukis, "API mashups: How well do they support the travellers' information needs?," *Procedia Comput Sci*, vol. 109, pp. 204–209, Jan. 2017, doi: 10.1016/J.PROCS.2017.05.326.

[2] Y. Xu, Z. Wang, H. Gao, Z. Jiang, Y. Yin, and R. Li, "Towards machine-learning-driven effective mashup recommendations from big data in mobile networks and the Internet-of-Things," *Digital Communications and Networks*, vol. 9, no. 1, pp. 138–145, Feb. 2023, doi: 10.1016/J.DCAN.2022.12.009.

[3] H. Wu, Y. Duan, K. Yue, and L. Zhang, "Mashup-Oriented Web API Recommendation via Multi-Model Fusion and Multi-Task Learning," *IEEE Trans Serv Comput*, vol. 15, no. 6, pp. 3330–3343, Nov. 2022, doi: 10.1109/TSC.2021.3098756.

[4] L. Yao, X. Wang, Q. Z. Sheng, B. Benatallah, and C. Huang, "Mashup Recommendation by Regularizing Matrix Factorization with API Co-Invocations," *IEEE Trans Serv Comput*, vol. 14, no. 2, pp. 502–515, Mar. 2021, doi:

10.1109/TSC.2018.2803171.
[5] "YandexGPT API," *Yandex Cloud - Documentation*. Accessed: Apr. 07, 2024. [Online]. Available: https://yandex.cloud/ru/docs/yandexgpt/
[6] M. Bilal, Y. Zhang, S. Cai, U. Akram, and A. Halibas, "Artificial intelligence is the magic wand making customer-centric a reality! An investigation into the relationship between consumer purchase intention and consumer engagement through affective attachment," *Journal of Retailing and Consumer Services*, vol. 77, Iss. 103674, p. 1–11. Mar. 2024, doi: 10.1016/J.JRETCONSER.2023.103674.
[7] L. Rudikowa, O. Myslivec, S. Sobolevsky, A. Nenko, and I. Savenkov, "The development of a data collection and analysis system based on social network users' data," *Procedia Comput Sci*, vol. 156, pp. 194–203, Jan. 2019, doi: 10.1016/J.PROCS.2019.08.195.
[8] S. Lamrhari, H. El Ghazi, M. Oubrich, and A. El Faker, "A social CRM analytic framework for improving customer retention, acquisition, and conversion," *Technol Forecast Soc Change*, vol. 174, p. 121275, Jan. 2022, doi: 10.1016/J.TECHFORE.2021.121275.
[9] A. K. Shahade, K. H. Walse, V. M. Thakare, and M. Atique, "Multi-lingual opinion mining for social media discourses: an approach using deep learning based hybrid fine-tuned smith algorithm with adam optimizer," *International Journal of Information Management Data Insights*, vol. 3, no. 2, p. 100182, Nov. 2023, doi: 10.1016/J.JJIMEI.2023.100182.
[10] E. Shchepina and A. Surikov, "Modeling the trajectories of interests and preferences of users in digital social systems," *Procedia Comput Sci*, vol. 212, no. C, pp. 104–113, Jan. 2022, doi: 10.1016/J.PROCS.2022.10.212.
[11] L. Miao, X. Li, D. Yu, Y. Ren, Y. Huang, and S. Cao, "Integrating users' long-term and short-term interests with knowledge graph to improve restaurant recommendation," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 9, p. 101735, Oct. 2023, doi: 10.1016/J.JKSUCI.2023.101735.
[12] Z. Shokrzadeh, M. R. Feizi-Derakhshi, M. A. Balafar, and J. Bagherzadeh Mohasefi, "Knowledge graph-based recommendation system enhanced by neural collaborative filtering and knowledge graph embedding," *Ain Shams Engineering Journal*, vol. 15, no. 1, p. 102263, Jan. 2024, doi: 10.1016/J.ASEJ.2023.102263.
[13] X. Wang et al., "Intent-aware Graph NN for Point-of-Interest embedding and recommendation," *Neurocomputing*, vol. 557, p. 126734, Nov. 2023, doi: 10.1016/J.NEUCOM.2023.126734.
[14] D. M. Obolensky, V. Shevchenko, O. Chengar, E. N. Maschenko, and A. S. Soina, "An implementation of social network group classification model based on recurrent and convolution NNs," *Economics. Information technologies*, vol. 48, no. 1, pp. 100–115, Sep. 2022, doi: 10.52575/2687-0932-2021-48-1-100-115.
[15] X. Song, J. Li, T. Cai, S. Yang, T. Yang, and C. Liu, "A survey on deep learning based knowledge tracing," *Knowl Based Syst*, vol. 258, p. 110036, Dec. 2022, doi: 10.1016/J.KNOSYS.2022.110036.
[16] T. K. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Comput Sci Rev*, vol. 40, p. 100395, May 2021, doi: 10.1016/J.COSREV.2021.100395.
[17] N. Ameen, A. Tarhini, A. Reppel, and A. Anand, "Customer experiences in the age of artificial intelligence," *Comput Human Behav*, vol. 114, p. 106548, Jan. 2021, doi: 10.1016/J.CHB.2020.106548.
[18] A. Micu, A. Capatina, D. S. Cristea, D. Munteanu, A. E. Micu, and D. A. Sarpe, "Assessing an on-site customer profiling and hyper-personalization system prototype based on a deep learning approach," *Technol Forecast Soc Change*, vol. 174, p. 121289, Jan. 2022, doi: 10.1016/J.TECHFORE.2021.121289.
[19] A. M. Tjoa, P. Wetz, E. Kiesling, T. D. Trinh, and B. L. Do, "Integrating Streaming Data into Semantic Mashups," *Procedia Comput Sci*, vol. 72, pp. 1–4, Jan. 2015, doi: 10.1016/J.PROCS.2015.12.098.
[20] G. Wang, J. Yu, M. Nguyen, Y. Zhang, S. Yongchareon, and Y. Han, "Motif-based graph attentional NN for web service recommendation," *Knowl Based Syst*, vol. 269, p. 110512, Jun. 2023, doi: 10.1016/J.KNOSYS.2023.110512.
[21] A. Capatina, M. Kachour, J. Lichy, A. Micu, A. E. Micu, and F. Codignola, "Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations," *Technol Forecast Soc Change*, vol. 151, p. 119794, Feb. 2020, doi: 10.1016/J.TECHFORE.2019.119794.
[22] L. Guo, R. Sharma, L. Yin, R. Lu, and K. Rong, "Automated competitor analysis using big data analytics: Evidence from the fitness mobile app business," *Business Process Management Journal*, vol. 23, no. 3, pp. 735–762, 2017, doi: 10.1108/BPMJ-05-2015-0065/FULL/XML.
[23] Y. Lee, "Extraction of Competitive Factors in a Competitor Analysis Using an Explainable NN," *Neural Process Lett*, vol. 53, no. 3, pp. 1979–1994, Jun. 2021, doi: 10.1007/S11063-021-10499-6/TABLES/8.
[24] "TargetHunter." Accessed: Apr. 07, 2024. [Online]. Available: https://targethunter.ru/
[25] "Livedune." Accessed: Apr. 07, 2024. [Online]. Available: https://livedune.com/ru/vk-analytics
[26] "Description of VK API methods for developers." Accessed: Apr. 07, 2024. [Online]. Available: https://dev.vk.com/ru/method