AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 58(2) (2026) 243-276
DOI: 10.22060/miscj.2026.25707.5487

Labeling Text Reviews for Customer Churn Prediction via Active Learning, Semi-Supervised Learning, and Explainable AI

Soroush Bajoghli, Fakhroddin Noorbehbahani*

Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.

ABSTRACT: Customer churn prediction is critical for businesses, yet most prior work relies on structured data and uses textual reviews sparingly, leaving low-resource languages such as Persian underexplored. Text-based churn models require large labeled corpora, which are often expensive and time-consuming to produce. To address this, we propose a practical method that combines least-confidence active learning, explainable AI (LIME) to assist expert annotators, and iterative semi-supervised labeling (self-training, label propagation, and a hybrid method). To our best knowledge, there is no study on using active learning, semi-supervised learning, and explainable AI to label textual data for churn prediction. Beginning with an expert-verified seed of 2,004 balanced Persian reviews, we trained LSTM/GRU classifiers on ParsBERT embeddings, augmented training sets via active sampling (100 samples per fold), and expanded labels to 10,500 Digikala reviews over 12 semi-supervised stages. Active learning increased average accuracy from 82.5% to 84.7%, and self-training produced the best sustained gains (mean peak accuracies of 86%), outperforming label propagation and the hybrid approach. Moreover, the explanations generated by LIME were positively evaluated by three domain experts, with mean ratings exceeding 4 for two experts and above 3.9 for the third, reflecting consistently high satisfaction. In addition, expert agreement increased by up to 8%, reaching 0.91 after reviewing the explanations. The proposed framework offers a reproducible solution for churn prediction in low-resource text domains by introducing the first labeled Persian review dataset for churn prediction and a unified framework integrating active learning, semi-supervised learning, and explainable AI.

Review History:

Received: Apr. 20, 2026
Revised: Jun. 30, 2026
Accepted: Aug. 11, 2026
Available Online: Sep. 15, 2026

Keywords:

Customer Churn Prediction
Labeling Persian Reviews
Active Learning
Explainable AI
Semi-supervised Learning

1- Introduction

Customer churn refers to customers discontinuing their purchase and use of a company's products or services, often switching to a competitor. Customer churn is a key topic in customer relationship management (CRM) and marketing. With saturated markets and increasing competition, particularly through social media, retaining customers and preventing churn has become a critical focus for businesses [1]. Given the high cost of acquiring new customers, organizations are increasingly prioritizing customer retention strategies over customer acquisition. Building valuable relationships with customers not only fosters loyalty but also leads to higher revenue and profitability compared to acquiring new customers.

Moreover, customer loyalty tends to strengthen over time, and loyal customers often refer new ones through word-of-mouth marketing. Studies have shown that increasing customer retention rates by just 5% can boost profits by up to 95% [2]. Additionally, predicting customer churn enables companies to allocate resources more effectively by focusing on at-risk customers. This targeted approach reduces unnecessary costs associated with acquiring new customers and improves return on investment. Therefore, customer churn prediction has become a vital concern in today's competitive business landscape [3].

Traditional approaches to customer churn prediction often rely on structured data such as demographic information, purchase history, and transactional records. While these methods may yield acceptable results, they frequently lack the capacity to provide clear explanations for why customers become disengaged or the underlying reasons behind their behavior. Moreover, correlation analyses between structured features and churn labels often reveal weak relationships, suggesting that relying solely on structured features may not guarantee optimal performance in churn prediction [4]. With the rapid growth of social media platforms, the number of customers actively engaging on networks such as blogs, Twitter, Facebook, and others has significantly increased. These platforms have become outlets for individuals to share their opinions, experiences, expectations, and demands. As a result, unstructured data has emerged as a valuable source for decision-making in identifying churn and detecting customer

*Corresponding author's email: noorbehbahani@eng.ui.ac.ir

Copyrights for this article are retained by the author(s) with publishing rights granted to Amirkabir University Press. The content of this article is subject to the terms and conditions of the Creative Commons Attribution 4.0 International (CC-BY-NC 4.0) License. For more information, please visit https://www.creativecommons.org/licenses/by-nc/4.0/legalcode.

UNCORRECTED PROOF

dissatisfaction with a company or product. Not only does this content heavily influence the purchasing decisions of other consumers, but it also provides marketers with deeper insights into customer behavior and preferences [5].

Customer churn prediction using textual analysis has received relatively limited attention in the literature, primarily due to the scarcity of suitable labeled datasets. This issue is particularly pronounced for low-resource languages such as Persian, where labeled datasets for churn prediction via customer reviews are virtually non-existent. Labeling customer reviews is also costly and labor-intensive. To mitigate this, active learning and semi-supervised learning techniques have shown promise by reducing annotation burden through selective sampling and by leveraging both labeled and unlabeled data [6]. To the best of our knowledge, however, these methods have not yet been applied to customer churn prediction from textual data, highlighting a clear research gap.

Among active learning strategies, uncertainty sampling is particularly effective in identifying the most informative samples for labeling, thereby improving model performance with fewer annotations [7]. Nevertheless, a key challenge of applying this approach with deep learning models is that these models often operate as "black boxes", leaving domain experts unable to understand why the model is uncertain about specific samples. Explainable AI (XAI) techniques such as LIME or SHAP can provide interpretable insights into model uncertainties and decision boundaries, thereby aiding expert judgment during labeling. The effectiveness of XAI has been well established in domains like healthcare for improving transparency and trust [8]. Despite these advantages, the integration of XAI with active learning for churn prediction remains unexplored. In this study, we bridge this gap by developing a novel pipeline that combines least-confidence active learning, semi-supervised self-training, and XAI, applied to Persian customer reviews. This approach enables cost-effective annotation, enhanced interpretability, and robust churn prediction in a resource-limited language setting.

This study addresses the following research questions:

- **RQ1:** To what extent is active learning effective in improving customer churn prediction performance?
- **RQ2:** To what extent do experts perceive the explanations generated by XAI as useful and effective in enhancing labeling agreement?
- **RQ3:** Does the use of semi-supervised learning improve evaluation metrics for customer churn prediction?
- **RQ4:** Which semi-supervised learning method performs best for labeling textual data in customer churn prediction?

This paper is organized as follows. Section 2 reviews prior work on customer churn from text, active learning, and semi-supervised methods. Section 3 describes our proposed labeling pipeline in detail: seed data creation, expert annotation and agreement, preprocessing, the LSTM/GRU churn models and hyperparameter tuning, the active-learning loop augmented with LIME explanations for annotators, and the three semi-supervised labeling strategies used in iterative stages under both randomized and balanced splits. Section 4 presents the results: frequent-term analyses by class, LIME example explanations and expert usefulness ratings, the impact of active learning on model metrics, and the stage-by-stage performance of the three semi-supervised methods. Finally, Section 5 summarizes the contributions, discusses limitations, and outlines directions for future work.

2- Related Work

This section reviews previous research relevant to the topic of this study. The primary focus is on customer churn prediction using textual data, as well as the application of active learning and semi-supervised learning approaches in this context. This section provides a comprehensive review of previous studies and is categorized into three main areas: Churn prediction from textual data, Active learning for textual data, and Semi-supervised learning for textual data.

2- 1- Churn Prediction from Textual Data

In recent years, various aspects of business have been explored to predict customer churn, offering valuable insights into customer behavior. However, most of the state-of-the-art frameworks currently focus exclusively on the use of structured data [10]. Existing churn prediction approaches have been widely applied in the telecommunications sector, where the increasing number of service providers has made churn prediction a critical necessity. This section reviews existing studies specifically related to customer churn prediction based on textual data and highlights the progress made in this area.

In recent years, content extraction combined with natural language processing on unstructured data has advanced customer churn prediction models and shown highly promising results. For example, sentiment analysis and emotional feature extraction have been used to assess the impact of emotional characteristics on churn in order to improve customer retention. Techniques such as logistic regression, support vector machines, and random forests have been applied using emotional features derived from customer-generated content (e.g., complaint and request emails) along with RFM features (Recency, Frequency, Monetary), resulting in a relatively high accuracy of 0.76 for churn prediction [11].

However, in many cases, emotional features and sentiment analysis alone do not provide sufficient insight into customer churn behavior. Amiri and Daume [12] proposed a method for identifying whether a customer is likely to churn in relation to a specific brand by analyzing their textual reviews. They developed a system to predict churn in microblogs using machine learning models such as support vector machines and logistic regression. The study also highlights the complexity of text classification for churn, emphasizing the subtle nature of churn-related content. Several linguistic challenges were noted for the underperformance of rule-based methods, such as: the comparison of competing brands, syntactic elements like "from" and "to" that may reverse the sentence meaning, the presence of negations, and misleading keywords (e.g.,

UNCORRECTED PROOF

"leaving") that do not necessarily indicate churn. The researchers examined three types of features: demographic features (derived from user profiles), content features (semantic information extracted from tweets), and contextual features (topics discussed between users or in relation to target or competing brands). The results showed that combining all three types yielded the best performance. In a follow-up study by the same authors [13], recurrent neural networks were used to model word embeddings and represent posts in a vector space for churn prediction. A dependency path approach was used to identify important attributes describing brands by including all words in the tweet and their syntactic relations with the mentioned brand. Furthermore, the study reiterated that churn classification differs significantly from sentiment analysis, as previous advanced sentiment analysis systems failed to perform effectively in churn detection.

In general, the use of neural networks has become a prevalent approach in the analysis of various types of textual data. Gridaj et al. [14] proposed a framework that integrates structured knowledge with unstructured knowledge derived from textual data, represented in the form of logical rules, to identify potential churn customers. Their model employs convolutional neural networks (CNNs) enhanced by injecting domain-specific logic rules, specifically, "but", "switch from", and "switch to", into the network's weights as prior knowledge. This hybrid method enables more accurate churn prediction by combining symbolic logic with neural representation learning. Additionally, they applied weight normalization to accelerate model training. Their approach was tested on the Twitter dataset previously used by Amiri and Daume [12], and showed improved classification performance over Amiri's two prior works. In a related study, it was demonstrated that combining different types of neural network modules, such as recurrent neural networks (RNNs), CNNs, regularization techniques, and attention mechanisms, can significantly enhance feature extraction, modeling capabilities, and weight assignment in text classification tasks [15]. The researchers conducted experiments on various datasets for three distinct text classification tasks: "topic classification", "argument mining", and "churn detection". The results confirmed that hybrid architectures integrating these neural modules outperform standard baseline models in textual churn prediction and other tasks.

In addition to collecting textual data from social media platforms such as Twitter and Facebook, customer call transcripts have also been converted into textual datasets for churn prediction tasks. In a study by Zhong and Li [16], a convolutional neural network (CNN)-based model was developed to detect churn signals using daily transcriptions of phone calls made by customers. To enhance the modeling of sequential text data, recurrent neural networks (RNNs) incorporating Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRUs) were utilized. Their results demonstrated that these deep learning architectures are effective in identifying churn-related signals from textual input. Specifically, the combination of CNNs and RNNs provides robust capabilities in feature extraction, sequence modeling, and attention-based weighting, contributing to improved performance in text classification tasks related to churn prediction.

In another study focused on customer churn prediction using textual data, chatbot interactions were leveraged as a valuable data source [17]. This research demonstrates that customer churn can be effectively identified through chatbot conversations and introduces a classification architecture that outperforms existing approaches for churn detection in social media contexts. The study further shows that by employing bilingual word embeddings, a model trained on a combined English-German dataset can surpass the performance of monolingual approaches. The researchers present new datasets in both English and German for churn detection, emphasizing the cross-lingual and cross-domain transferability of churn detection models. Key contributions of this study include a neural network-based architecture for churn classification, a multilingual approach utilizing cross-lingual word embeddings, the identification of churn-related patterns within chatbot dialogues, and the development of datasets derived from German tweets and bilingual chatbot conversations in English and German.

Although numerous studies have been conducted in the field of Persian text processing and sentiment analysis [18], there is a notable scarcity of research specifically focused on customer churn prediction using textual data in the Persian language. One notable study addressing this gap was carried out by Taherkhani et al. [19], which highlights the importance of churn prediction in the hospitality industry. This research focuses on the analysis of customer reviews written in Persian, utilizing text mining techniques and employing the Random Forest algorithm for classification tasks. The dataset used in this study was collected from customer databases available on social networks and hotel websites, with a specific focus on hotels located on Kish Island. To enhance the performance of their proposed model, the authors incorporated a gravitational search algorithm for feature selection and a differential evolution algorithm for hyperparameter optimization of the classifier. After performing preprocessing steps, parameter tuning, and removing irrelevant features, the study reported a significant improvement in model performance. Specifically, the metrics for precision, recall, accuracy, and F1-score were reported as 0.77, 0.76, 0.77, and 0.76, respectively.

Table 1 presents a summary of the reviewed studies in this field, along with the best reported F1-score.

2- 2- Active Learning for Textual Data

Text classification using active learning has been extensively studied as an effective approach for improving classification performance while minimizing the effort required for data annotation. Several studies have explored various active learning strategies and their applications across different text classification tasks. However, to the best of our knowledge, active learning has not yet been applied to unstructured data in the context of customer churn prediction. This section first reviews studies that focus on active learning for English-language text classification. Subsequently, the

UNCORRECTED PROOF

Table 1. Literature Review on Customer Churn Prediction for Textual Data. (Continued)

Study	Language Examined	ML Method	Dataset	Description	Limitations	Unresolved Innovation Gap	F1-score
[11]	English	Logistic Regression, SVM, and Random Forest	Customer subscription data from a newspaper	Utilizing customer sentiment features from their emails along with RFM features	Dataset limited to emails from a single company, restricting generalizability	No active learning or semi-supervised annotation framework; not designed for Persian textual churn prediction	0.76
[12]	English	Linear Classification, SVM, and Logistic Regression	Three textual datasets from telecommunication companies (Verizon, T-Mobile, and AT&T) collected from Twitter	Collecting and labeling customer review data from Twitter	Tweets are short with weak context; data manually annotated, costly and error-prone	Relies on fully manual labeling and English Twitter data; no scalable annotation strategy	0.80
[13	English	RNN, SVM, and Logistic Regression	Three textual datasets from telecommunication companies (Verizon, T-Mobile, and AT&T) collected from Twitter	Investigating challenging factors such as sentiment analysis and rule-based classifiers in churn prediction	Short-text representation struggles with informal language; manual labeling limits scalability	No active learning, semi-supervised learning, or explainability support for annotation	0.78
[14]	English	CNN combined with rule-based methods	Three textual datasets from telecommunication companies (Verizon, T-Mobile, and AT&T) collected from Twitter	Combining rule-based approaches with neural networks	Small and noisy dataset; manual annotation introduces dependency on annotator accuracy	Requires manual annotation and focuses on English microblogs only	0.83
[15]	English	RNN, CNN, regularization techniques, and attention mechanism	Three textual datasets from telecommunication companies (Verizon, T-Mobile, and AT&T) collected from Twitter	Integrating regularization methods and attention mechanisms with recurrent and convolutional neural networks	Hybrid CNN/RNN requires heavy hyperparameter tuning and uses datasets of limited relevance	Does not address annotation cost reduction or Persian-language churn prediction	0.87
[16]	English	RNN	Textual data from daily customer phone calls in the telecommunications industry	Feature extraction from text	Analysis based only on one telecom operator's call transcripts, limiting applicability	Lacks scalable data-labeling mechanisms and explainable annotation support	0.86

UNCORRECTED PROOF

Table 1. Literature Review on Customer Churn Prediction for Textual Data.

[17]	English-German	CNN and GRU	Data collected from Twitter and chatbots in English and German	Using chatbots to detect customer churn in both English and German	Multilingual coverage remains limited; manual labeling adds cost and risk of inconsistency	No integration of active learning and semi-supervised learning for dataset construction	0.85
[19]	Persian	Random Forest	Customer reviews collected from social networks and hotel websites	Applying the gravitational search algorithm for feature selection in Persian	Applied only to hotel industry data from one region; dataset is small, and the labeling process is unclear	Does not provide a reusable annotation framework; limited to a small Persian hotel dataset	0.76

role of active learning in Persian-language texts will be examined.

In a study conducted by Zhu et al. [20], active learning is explored in the context of word sense disambiguation and text classification. The authors introduce a novel selective sampling method that employs a k-nearest neighbors-based density measure to determine whether an unlabeled instance is relevant or not. In addition, they propose a method to construct an initial training dataset for active learning. The study presents an algorithm that combines these two sampling strategies within an active learning framework. The authors emphasize that their proposed methods are straightforward to implement and can be readily applied to various classification models, including Maximum Entropy, Naïve Bayes, and Support Vector Machines (SVMs).

In a subsequent study, a novel approach is proposed to reduce the effort and cost of labeling textual data for multi-label text classification tasks. Traditional active learning algorithms are typically designed for single-label problems, where each data instance is assigned only one label. However, in multi-label classification, an instance may be associated with multiple labels simultaneously. The authors propose an active learning algorithm that explicitly considers the multi-label nature of the data, aiming to select instances for labeling that are expected to yield the greatest reduction in classification error. The approach utilizes Support Vector Machines (SVMs) to estimate the model's prediction error and incorporates a strategy for predicting probable labels for unlabeled data. This method is evaluated across seven benchmark datasets and demonstrates promising results, achieving competitive performance while requiring significantly fewer labeled instances compared to state-of-the-art techniques [21].

Despite the valuable contributions of the two aforementioned studies to the fields of text classification and active learning, they did not address the issue of class imbalance, which was first highlighted by Li et al. In their research, the authors focused on active learning for sentiment analysis with imbalanced data and introduced a novel active learning framework that dealt with class imbalance by considering uncertainty in both positive and negative sentiment samples. Their approach effectively identified informative instances from both classes, resulting in improved classification performance, particularly on imbalanced datasets [22].

Subsequently, with the rise of deep neural networks, Zhang et al. were among the first to combine active learning with deep learning for text classification. They claimed to be the first to implement active learning in conjunction with Convolutional Neural Networks (CNNs) for this task. The study proposed a query strategy that selects samples based on the expected change in word embeddings and model uncertainty, thus enabling the model to learn more discriminative embeddings for text classification. However, although the paper reports improved performance over baseline methods, it lacks detailed evaluation metrics and a comprehensive comparison with other advanced approaches. This omission makes it difficult to fully assess the effectiveness and superiority of the proposed method. Nevertheless, due to its innovative integration of active learning and deep neural networks, the study remains a noteworthy contribution to the field [23].

In another study, an innovative approach to text classification using active learning combined with Support Vector Machines (SVMs) was proposed. Although, unlike the previously mentioned work, this study did not employ deep neural networks, it nonetheless achieved comparable and, in some cases, even superior results. The proposed method adopts a pool-based active learning strategy, in which informative samples are selected from a pool of unlabeled data based on the posterior probabilities generated by a set of multi-class SVM classifiers. These selected samples are then manually labeled by an expert. The primary objective of this approach is to guide the sample selection process in such

UNCORRECTED PROOF

a way that only a minimal number of documents need to be labeled before achieving a satisfactory level of classification accuracy [24].

To date, limited yet valuable research has been conducted on text classification using active learning for Persian texts. However, two notable studies have made pioneering contributions in this field and achieved promising results.

Ghayoomi and Kuhn [25] investigated entropy-based sampling and query-by-committee for expanding the Persian treebank by identifying informative samples. The authors addressed the challenges associated with manual annotation of large syntactic trees, a process that is both time-consuming and costly. They proposed the use of active learning to optimize the annotation process by selecting the most informative samples, thereby reducing labeling effort while still maintaining a high-quality treebank.

In another study, Asli et al. [9] explored the challenges of obtaining labeled data for training deep learning models, particularly in low-resource domains or languages. The paper introduced active learning strategies in which the model selects the most informative samples for labeling to improve labeling efficiency. The authors proposed a novel active learning strategy for text classification tasks based on Latent Dirichlet Allocation (LDA). This method aims to capture a global representation of the textual data by clustering it into topics, and then employs entropy-based sampling to query the most informative samples within each topic cluster. The approach outperformed other uncertainty-based methods in terms of classification performance.

Although these two studies represent important progress in applying active learning to Persian texts, there are still significant gaps in this area. For instance, a wider range of datasets should be evaluated, and active learning methods need to be applied to tasks beyond sentiment analysis. Furthermore, increasing final model accuracy remains a challenge, one that this research aims to address.

Table 2 presents a summary of the reviewed studies in active learning, along with the best reported F1-score.

Table 2. Literature Review on Active Learning for Textual Data.

Study	Language Examined	ML Method	Active Learning Method	Domain	Unresolved Innovation Gap	F1-score
[20]	English	KNN	Combined uncertainty-and-density sampling (SUD-SBC)	General NLP for text and word sense disambiguation	Does not combine active learning with semi-supervised learning or churn prediction	0.87
[21]	English	SVM	Multi-label active learning optimizing expected loss reduction	General text corpora - news and multi-topic content	Focuses on generic text classification rather than churn prediction and dataset annotation	0.87
[22]	English	SVM	Co-selecting two complementary classifiers to prioritize minority-class samples	Social media sentiment analysis	No explainable AI support and no Persian-language evaluation	0.82
[23]	English	CNN	Representation-focused active learning based on expected embedding changes	General text classification (sentence & document level)	Does not address large-scale annotation through semi-supervised learning	0.90
[24]	English	SVM	Batch selection using multi-class SVMs	News articles and scientific texts	Lacks integration with explainable AI and churn prediction tasks	0.97
[25]	Persian	word- and class-based parsing	Combined uncertainty model and query-by-committee for parsers	Computational linguistics - treebank creation (Persian)	Applied to Persian treebank construction rather than customer-review annotation	0.56
[9]	Persian	RNN and CNN	LDA-based sampling with entropy for informative instance selection	Persian sentiment analysis (social media - product reviews)	Focuses on sentiment analysis; no churn prediction or semi-supervised labeling framework	0.78

UNCORRECTED PROOF

2- 3- Semi-Supervised Learning for Textual Data

This section reviews a selection of significant studies on text classification using semi-supervised learning for both English and Persian texts. One notable study in this area focuses on the sentiment classification of MOOC (Massive Open Online Courses) posts, which is crucial for instructors to understand student emotions and enhance learning outcomes. In this study, the authors propose a semi-supervised deep learning model based on co-training, which achieves performance comparable to models trained on large labeled datasets, even when only a limited amount of labeled data is available. The research demonstrates that by leveraging a small set of labeled samples along with a large pool of unlabeled data, the proposed model yields promising results, even outperforming models trained exclusively on larger labeled datasets. This approach offers valuable assistance to instructors in monitoring learner sentiments and conducting timely interventions to enhance learning outcomes in MOOCs [26].

In another study, adversarial networks have been employed for text classification, presenting an innovative approach in the field. Briefly, adversarial networks, commonly known as Generative Adversarial Networks (GANs), are a class of machine learning models composed of two neural networks: a generator and a discriminator. The core idea behind GANs is that these two networks engage in a zero-sum game, competing with one another. The authors address the limitations of existing methods, particularly those based on generative models, which inherently struggle with textual data. To overcome this, they formulate semi-supervised learning as a model-based reinforcement learning problem, introducing two networks: a prediction network for estimating targets, and a judgment network for evaluation. The judgment network generates reward signals that guide the training of the prediction network, which is optimized using policy gradient methods. They propose a recurrent neural network-based model for semi-supervised text classification and conduct experiments on real-world benchmark datasets, demonstrating that their approach outperforms several state-of-the-art methods. Additionally, the study provides a comprehensive review of related work in semi-supervised learning and discusses the similarities and distinctions between their proposed framework and existing approaches [27].

An important area in semi-supervised learning is the application of graph-based networks, which have been widely used in text classification, particularly for fake news detection. In this context, one study addresses the problem of identifying fake news on social media platforms. The authors highlight the challenges posed by the spread of misinformation and the limited availability of labeled data for training. To address these issues, they propose a semi-supervised learning approach based on Graph Neural Networks (GNNs). The proposed framework consists of three main components: word embedding of the news content, graph construction, and graph-based classification. The authors use word embeddings to represent the textual content of news articles in a lower-dimensional space, capturing textual similarities among the articles. Then, they construct a graph based on these embeddings to model the relationships between articles. Finally, they apply Graph Neural Networks to perform classification and detect fake news [28].

Another significant application of semi-supervised learning is in the classification of short texts. In this regard, a study addresses the challenges of short text classification and proposes a novel method called Heterogeneous Graph Attention Networks for semi-supervised short text classification. This method utilizes a heterogeneous information network framework to model short texts and integrate auxiliary information. A two-level attention mechanism is employed to learn the importance of neighboring nodes and different types of nodes during classification. Experimental results demonstrate that the proposed model outperforms state-of-the-art methods on benchmark datasets. The study also discusses related works in traditional text classification, deep neural networks, and semi-supervised text classification [29].

Another innovative study addresses the challenge of selecting an effective threshold in semi-supervised learning and introduces a novel approach called Binary-Classifier Filters for semi-supervised learning. The proposed method utilizes binary classifiers as data filters to label the unlabeled data. The paper presents three techniques for labeling the unlabeled samples: cascaded binary classifiers, non-cascaded binary classifiers, and rank-based binary classifiers. Through extensive experiments across various domains, including image classification, text classification, and audio classification, the study demonstrates that rank-based binary classifiers achieve the best performance. The proposed approach outperforms traditional supervised learning methods in several scenarios, particularly when the number of labeled samples is limited, showing notable improvements in such cases [30].

There are only a limited number of studies that have explored semi-supervised learning for Persian textual data. One of the most notable works in this area introduces a model called BERT-SAWS for early detection of Persian rumors. This model leverages content-based and contextual features to improve the accuracy of rumor verification, focusing specifically on contextual word embeddings, speech acts, and writing style features. By analyzing these characteristics, the model is better equipped to understand the linguistic nature of rumors and verify them more effectively. Experimental results show that BERT-SAWS improves classification performance by 2% to 18% compared to baseline models. The study highlights the significance of speech act features and writing style in the rumor verification task and discusses the negative impacts of rumor spread as well as the need for automated systems to detect and control rumors on online social media platforms. The proposed model aims to address this issue by analyzing the content and contextual information of rumors without relying on structural metadata. Overall, the study demonstrates the application of natural language processing (NLP) techniques in detecting and verifying Persian rumors, while also addressing the challenge of limited reliable

UNCORRECTED PROOF

datasets in low-resource languages such as Persian. To overcome this challenge, the model leverages a pre-trained BERT and fine-tunes it on a small dataset of Persian rumors. The study's contributions include the development of a semi-supervised learning model for rumor verification, the ability to verify rumors in the early stages of dissemination, and the validation of its effectiveness on Twitter and Telegram datasets. Experimental results confirm that the proposed model outperforms baseline models and achieves higher accuracy in Persian rumor verification [31].

Another topic that has been explored in the context of semi-supervised learning for the Persian language is word sense disambiguation (WSD). A study in this regard proposes a model aimed at providing a solution for semantic disambiguation and overcoming the challenges posed by the rich use of metaphor and simile in Persian. The proposed framework is designed for disambiguating ambiguous words in Persian texts and sentences and is extensible, allowing for the integration of new algorithms to improve results. The study also points out that most existing research on word sense disambiguation relies on supervised methods and the availability of large labeled datasets. However, due to the lack of such resources in Persian, the authors adopt a semi-supervised learning approach. Their method leverages defined features of target words and employs co-training to enhance learning. For evaluation, they use a reference corpus of news articles published by news agencies. The implemented approach demonstrates promising results in correctly identifying the meanings of ambiguous words in Persian, achieving high accuracy, coverage, and precision [32].

In general, text classification using semi-supervised methods has gained increasing importance due to its ability to reduce labeling costs while achieving valuable results [33]. Currently, six key directions have been identified for the development of semi-supervised learning that can be leveraged in future research: Neural network-based language models for vector representation of text, Algorithms for hyperparameter optimization, Explainable artificial intelligence (XAI), a set of techniques that help users better understand the outcomes produced by machine learning algorithms, Regularization techniques, Resource development for languages other than English, and Analysis of harmful factors in semi-supervised learning in relation to unlabeled samples [33].

Table 3 presents a summary of the reviewed studies in this field, along with the best reported F1-score.

Table 3. Literature Review on Semi-Supervised Learning for Textual Data.

Study	Language Examined	ML Method	Semi-Supervised Learning Method	Domain	Unresolved Innovation Gap	F1-score
[26]	English	RNN	Co-training	MOOC forum sentiment classification - online education	Does not incorporate active learning or explainable AI for sample selection	0.93
[27]	English	RNN	Adversarial networks with policy gradient	General text classification - NLP	Designed for generic text classification rather than churn prediction	0.90
[28]	English	CNN and KNN	Label propagation	Fake news detection - social media analysis	Applied to fake-news detection; lacks active learning-based annotation	0.86
[29]	English	CNN and RNN	Label propagation	Short-text NLP - social media and microblogs	Focuses on short-text classification and does not address annotation efficiency	0.82
[30]	English	PNN	Self-training	General text classification - information retrieval	No explainable AI component and no customer churn application	0.90
[31]	Persian	Random Forest, SVM and DNN	Self-training	Persian rumor verification - social media	Applied to rumor verification; lacks active learning integration	0.93
[32]	Persian	Decision List Method	Self-training	Persian word sense disambiguation - NLP	Does not address customer churn prediction or explainable annotation support	0.96

UNCORRECTED PROOF

2- 4- Summary

Prior studies on text classification have shown that active learning and semi-supervised learning can effectively reduce annotation costs while improving predictive performance. Active learning often relies on uncertainty-based strategies such as least-confidence sampling to select the most informative examples for labeling, whereas semi-supervised methods including self-training and label propagation approaches leverage unlabeled data to enhance model generalization. These methods have been successfully applied in various text domains; however, to date, they have not been employed for customer churn prediction, especially using unstructured Persian text.

As shown in Tables 1–3, previous studies have addressed customer churn prediction, active learning, and semi-supervised learning independently. However, several important research gaps remain unresolved. First, none of the reviewed churn prediction studies proposed a scalable annotation framework that combines active learning and semi-supervised learning to reduce manual labeling costs. Second, existing active learning studies primarily focused on generic text-classification tasks and did not investigate Persian customer churn prediction. Third, previous semi-supervised learning studies did not integrate explainable AI techniques to support expert-driven annotation. The proposed method addresses these gaps by introducing a unified framework that combines active learning, explainable AI, and semi-supervised learning for Persian customer churn prediction. Furthermore, the study provides the first labeled Persian dataset for textual churn prediction, thereby contributing a new resource for future research in this domain.

Despite the growing body of research on text-based customer churn prediction, most existing studies have focused on English datasets and the telecommunications domain. In contrast, research on Persian textual data remains very limited, and the few available studies are mainly concerned with sentiment analysis or other application domains rather than customer churn prediction. In particular, the use of customer reviews from Persian e-commerce platforms, such as Digikala, has received little attention. To the best of our knowledge, no publicly available labeled dataset of Persian online store reviews exists for churn prediction, and the combination of active learning, semi-supervised learning, and explainable AI for efficient annotation of such data has not been investigated. This gap motivates the present study and highlights the need for developing scalable labeling approaches for low-resource Persian e-commerce environments.

3- Method

The proposed method for labeling churn reviews is illustrated in Figure 1. According to Figure 1, the process begins with seed data creation. The reference dataset in this study consists of user reviews collected from the Digikala website [34]. As the largest e-commerce platform in Iran, Digikala hosts a substantial volume of customer reviews on its products, many of which are publicly available and unlabeled. The ultimate objective of this study is to label 10,500 of these reviews for the purpose of predicting customer churn, using a combination of active learning and semi-supervised learning. Each review is to be categorized into one of three defined classes: "*Not-Churn*", "*Unknown,*" and "*Churn*." The number 10,500 was chosen as a target due to its considerable size, balanced with practical constraints such as time and available resources. To achieve this goal, the first step involves seed data creation. For this purpose, a subset of Digikala's reviews, a selection of data from Amiri's study [10], and a number of synthetically generated reviews by ChatGPT (GPT-4o) were selected. These reviews were then labeled by three domain experts. As a result, 2,004 labeled reviews were obtained, each assigned to one of the three classes ("*Not-Churn*", "*Unknown,*" and "*Churn*"), thus forming the seed labeled dataset for the study.

After obtaining the seed labeled dataset, the agreement criteria among the experts are calculated using the Fleiss' Kappa coefficient. In the next step, the preprocessing of the reviews in the seed data is carried out, which includes tasks such as removing extra spaces, punctuation marks, numbers, and stop words, as well as performing normalization and word vectorization. Once the text cleaning and word embedding are complete, the dataset is prepared for evaluation using cross-validation. To this end, the seed dataset is divided into six folds, where in each iteration, one fold is used as the test set. It is worth noting that this folding process is conducted in two ways: randomized splitting and balanced splitting. In randomized splitting, reviews are randomly assigned to folds, which may result in an imbalanced distribution, such as having more "*Churn*" reviews in one fold compared to the other two categories. In contrast, the balanced splitting ensures an equal number of reviews from each of the three classes in all six folds. Accordingly, the rest of the study is conducted in both scenarios: one with randomly distributed folds and another with balanced folds. The outcomes of these two settings will be compared in the results section.

In the next step, a customer churn prediction model based on LSTM and GRU is trained using the training samples (each time consisting of 5 out of the 6 folds). For evaluation, as previously mentioned, the cross-validation method is employed, where in each iteration, one fold is used as the test set and the remaining folds serve as the training set. As a result, six distinct training datasets are formed, each with a different test fold. The average performance across these six folds is then considered the final evaluation metrics for the initial labeled dataset.

Subsequently, active learning and explainable artificial intelligence (XAI) are applied. For each of the six initial training datasets, 100 additional reviews from the Digikala dataset, labeled by experts with the assistance of active learning and XAI, are added, and the models are re-evaluated on the same test sets to observe any performance improvements. In the following phase, semi-supervised learning is employed using three approaches: self-training, label propagation, and a hybrid method that combines both. The remaining unlabeled Digikala reviews are labeled using these methods and are

UNCORRECTED PROOF

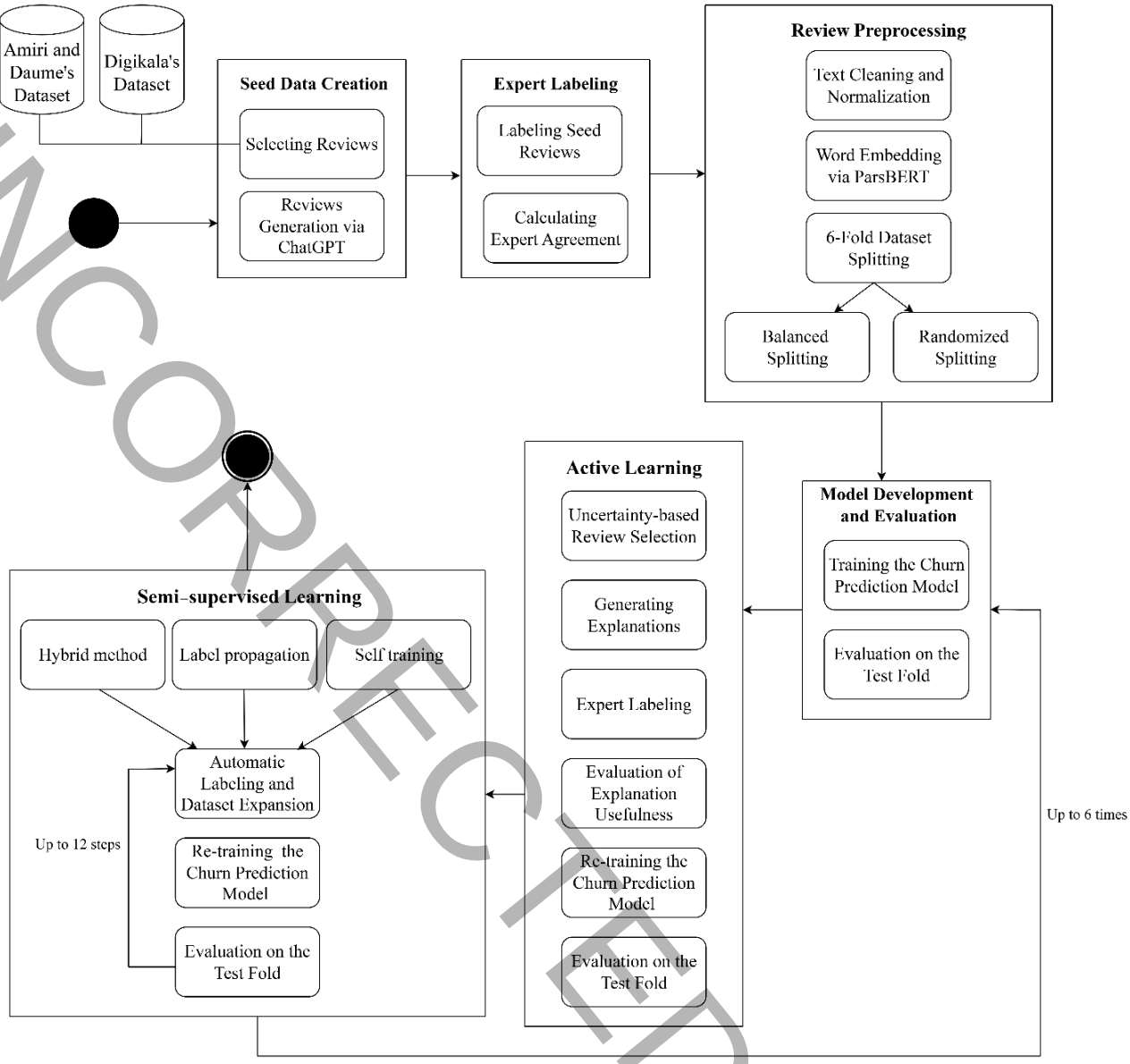


Fig. 1. Proposed Labeling Method.

gradually added to each of the six training datasets over 12 iterations. Evaluation metrics are again computed on the same initial test folds to examine the degree of improvement and to identify the most effective semi-supervised learning approach.

3- 1- Seed Data Creation

The first and most critical step in this study is the construction of the seed dataset. As illustrated in Figure 1, this process is carried out through two main sources:

- **Selecting Reviews**

As previously mentioned, a large corpus of customer reviews from the Digikala platform is publicly available. The goal of this study is to label 10,500 of these reviews to support customer churn prediction. For the seed data, 1,250 reviews were selected to be annotated by three domain experts. The remaining 9,250 samples were later labeled using active learning and semi-supervised learning techniques.

The next portion of the seed data was derived from the labeled dataset used in the study by Amiri et al. [10], which was originally compiled for customer churn prediction in the telecommunications sector. A total of 300 English reviews from this dataset were translated into Persian and included in the seed data. The rationale for using this dataset was to ensure comprehensive coverage of diverse churn-related expressions in the seed data, including phrases such as *"leaving the company"* or *"subscription cancellation."*

UNCORRECTED PROOF

However, not all samples from Amiri's dataset were used: many of the *Not-churn* samples were excluded to maintain class balance, and some *Churn* samples were discarded due to redundancy or lack of informative content (e.g., samples with only one word). Although the original dataset included labels, all 300 translated reviews were re-annotated by the same three experts to ensure consistency and labeling accuracy.

• Reviews Generation via ChatGPT

Another portion of the seed data was generated using ChatGPT. Initially, the model was prompted to generate Persian-language customer reviews across all three target classes: "*Not-Churn*", "*Unknown,*" and "*Churn*". After generating the primary reviews, ChatGPT was further prompted to create paraphrases and semantically similar reviews to augment the dataset. This process was also repeated in English, and the resulting outputs were translated into Persian, based on the observation that the English-generated samples were often more coherent and semantically relevant. In total, 454 synthetic reviews were created through this method.

Examples of prompts provided to ChatGPT include:

- Write a 1–2 sentence customer review (1 star) for [BRAND] saying the customer will cancel immediately, asking for a refund, and naming the main reason (e.g., billing error, broken product, or poor service).
- Write a 2–4 sentence review (2 stars) in a calm tone for [BRAND]. Describe repeated problems and failed support attempts, and end with a clear statement that the customer will stop using the product/service.
- Write a short review (1–3 sentences) for [BRAND] saying the customer is switching to [COMPETITOR] because of price/features/quality. Include one concrete comparison (e.g., "cheaper plan" or "better customer service").
- Compose a 2–3 sentence review for a subscription service [BRAND]. Mention an unexpected auto-renewal or recurring charge and the customer's decision to cancel the subscription immediately.
- Write a review (3–5 sentences) describing a defective product from [BRAND], the return/repair experience, and concluding that the customer will not buy from the brand again.
- Write a short review (1–3 sentences) for [BRAND] about a late/missing delivery for an urgent order and state that, because of this, the customer will no longer purchase from the brand.
- Write a review (2–4 sentences) where [BRAND] raises prices or changes terms. The customer explains why the new policy is unacceptable and that they will cancel/leave.
- Write a single-sentence, tweet-style review (≤280 characters) announcing that you've canceled [BRAND] and briefly why (e.g., "switched to X", "no support", "charged me twice").

After generation, all reviews were evaluated by the three domain experts, and only those unanimously considered semantically relevant and contextually suitable for churn prediction were included in the final dataset.

In total, 2,004 labeled samples were compiled for the initial dataset. The distribution of samples across the three classes is presented in Table 4.

3- 2- Expert Labeling

The second step in this study is expert labeling. As illustrated in Figure 1, this process consists of two parts:

• Labeling Seed Reviews

In this phase, the reviews of the seed data were annotated by three domain experts. As previously described, these reviews were obtained from three sources: *Digikala customer reviews*, *ChatGPT-generated content*, and the *Amiri et al. study*. The selection of reviews for annotation was carried out by one of the three experts. In each iteration, this expert selected a set of samples and annotated them, then passed them to the remaining two experts for independent annotation. The final label for each sample was determined after collecting the annotations from all three experts. This iterative process continued until a total of 2,004 labeled samples were obtained. The main reason for adopting a stepwise approach was to ensure class balance in the initial dataset. For example, if at a certain stage 200 samples had been labeled and the *Unknown* class contained fewer samples than the other two classes (*Churn* and *Not-Churn*), the expert responsible for sample selection would deliberately choose new samples that, in their judgment, were more likely to

Table 4. Distribution of samples constituting the initial dataset.

Dataset	Number of Samples	Percentage
Digikala	1250	62.4%
ChatGPT	454	22.6%
Amiri et al.	300	15.0%

UNCORRECTED PROOF

belong to the *Unknown* category.

This balancing process was not without challenges. Naturally, in some cases, the judgments of the other two experts differed from those of the selecting expert, which could lead to minor imbalances. However, through discussion and consensus among all three experts, and, in some cases, the removal or addition of certain samples, class balance was ultimately achieved. The final labeled dataset contained exactly 668 samples in each of the three categories (*Churn*, *Unknown*, and *Not-Churn*), ensuring a perfectly balanced distribution for the initial phase of model training.

In the constructed dataset, samples in which the user explicitly stated that they would no longer purchase from the targeted brand or company were assigned the label *Churn*. Samples where the user merely expressed an opinion about a product, often indicating dissatisfaction, but did not explicitly mention whether they would make future purchases were assigned the label *Unknown*. Conversely, reviews where the user expressed strong satisfaction with the product were labeled as *Not-Churn*. It should be noted that although customer churn can generally be identified over time through behavioral tracking, the primary aim of this study was to leverage explicit customer statements indicating future disengagement from the brand. By detecting such instances of churn intention early, proactive measures can be taken to prevent customer loss. Accordingly, these explicit churn-indicative reviews were labeled as *Churn* with the assistance of three domain experts.

• Calculating Expert Agreement

To assess the consistency among the experts involved in the annotation process, the inter-annotator agreement was calculated using Fleiss' Kappa coefficient. This statistical measure evaluates the degree of agreement among multiple raters assigning categorical labels to a fixed set of instances, while accounting for the agreement occurring by chance. In this study, the experts classified each customer review into one of three categories: Churn, Unknown, or Not churn. The obtained Fleiss' Kappa score was used to validate the reliability of the labeling process and ensure the consistency of expert judgments before proceeding to the subsequent data preprocessing and modeling stages.

3- 3- Review Preprocessing

This part describes the preprocessing steps applied to the seed data. In general, the preprocessing pipeline consists of three main operations:

• Text Cleaning and Normalization

Initially, consecutive and unnecessary spaces are eliminated to standardize word spacing. This step facilitates proper tokenization and ensures consistency in the textual data. In addition, punctuation marks and numerical values are removed, as they usually carry no significant semantic meaning in this context and cannot be effectively represented during vectorization. Removing such elements reduces storage requirements and simplifies subsequent processing.

In the normalization process, all texts are standardized to ensure consistency. This includes the removal of orthographic symbols such as tashdid (diacritic for emphasis) and the two dots below the letter "ی," which occasionally appear in Persian words. Furthermore, stop words are also removed. Stop words are frequently used words in a language (e.g., "به", "از", "و" in Persian) that often carry little meaningful information. These words typically hold little significance in conveying the sentence's meaning, and their removal does not affect the sentence's comprehension [35]. Eliminating these words helps reduce noise and enhances the performance of text processing algorithms.

• Word Embedding via ParsBERT

In the next preprocessing step, textual data are transformed into vector representations using the ParsBERT language model. Since ParsBERT generates rich contextualized embeddings for all tokens in a sentence, two sentence representation strategies were evaluated in this study: mean pooling over all token embeddings and the embedding corresponding to the special [CLS] token. Experimental results showed that the two strategies produced comparable performance across all evaluation metrics. This can be attributed to the strong contextual representations learned by ParsBERT during pre-training, which encode semantic information across all token embeddings. Consequently, both mean pooling and the [CLS] representation are capable of capturing the overall semantics of customer reviews effectively, resulting in only marginal performance differences in the downstream customer churn prediction task. Although mean pooling achieved a slightly higher accuracy (approximately 0.3%) than the [CLS] representation, the difference was negligible. Therefore, due to its marginally better empirical performance, mean pooling was adopted as the sentence representation method, and all experimental results reported in the subsequent sections are based on this representation.

• 6-Fold Dataset Splitting

As the third step of the preprocessing phase, the processed reviews were partitioned to prepare the data for subsequent model evaluation. A 6-fold cross-validation strategy was employed, in which the dataset was divided into six folds, with one fold used as the test set and the remaining folds used for training in each iteration. This folding process was carried out in two configurations: randomized splitting and balanced splitting. In the randomized splitting, instances were randomly assigned to folds, which could lead to uneven class distributions, such as having more "Churn" samples in one fold compared to the other two categories. Conversely, the balanced splitting maintained an equal number of samples from each class across all folds. Both configurations were retained for later stages to enable a comparative analysis of model performance under different data distribution conditions.

3- 4- Model Development and Evaluation

This stage focuses on building and assessing the customer

UNCORRECTED PROOF

churn prediction model based on the prepared dataset. It is divided into two main components:

• Training the Churn Prediction Model

The next step involves building and training the customer churn prediction model. Due to their high performance and representational power, deep neural networks were chosen as the primary modeling approach. Specifically, two widely recognized and effective architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, were utilized in this study. A critical aspect of constructing a deep learning model is the selection of appropriate hyperparameters for different layers. To address this, the keras_tuner library was employed to automatically explore different hyperparameter configurations. Specifically, the Hyperband algorithm was used with validation accuracy as the optimization objective. A validation subset comprising 20% of the data was used exclusively for hyperparameter tuning, while the final model evaluation was performed using six-fold cross-validation. The search space included the number of units in the LSTM layers, the GRU layer, and the dense layer, as well as the corresponding dropout rates. Table 5 summarizes the search ranges and the final values selected by the tuning process. The Adam optimizer with its default initial learning rate was used, and an exponential learning rate decay schedule was applied after the tenth epoch. This strategy ensured a systematic and reproducible selection of hyperparameters, thereby enhancing the reliability of the developed models.

The architecture of the proposed model is illustrated in Figure 2.

• Evaluation on the Test Fold

In this stage, performance assessment was carried out using a 6-fold cross-validation procedure. In each iteration, one fold was designated as the test set, while the remaining five folds were used for training. The trained model was then evaluated on the test fold, and performance metrics including accuracy, precision, recall, and F1-score were computed. This process was repeated six times to ensure robust and unbiased performance estimation. Finally, results obtained from the randomized and balanced data configurations were compared to examine the influence of data distribution on model stability and predictive performance.

3- 5- Active Learning

The Active Learning phase is designed to iteratively enhance the model's performance by incorporating expert feedback on the most informative samples. At this stage, following the manual labeling of 1,250 out of the total 10,500 Digikala reviews, there remain 9,250 unlabeled reviews. The ultimate goal is to label all these 9,250 reviews in the context of customer churn prediction. As previously mentioned, six distinct models have been trained (based on 6-fold cross-validation), and each of these models will now be used to select 100 reviews for expert annotation via active learning. That is, the first model selects 100 reviews from the pool of 9,250 for labeling, then the second model selects another 100, and so on until all six models have performed their selections. It should be noted that the selected reviews may overlap among models, meaning that some samples might be chosen by more than one model. Therefore, the result is not necessarily 600 unique reviews, but rather up to 600 reviews

Table 5. Hyperparameter Search Space and Optimal Values Selected by Keras Tuner.

Hyperparameter	Search Range	Selected Value
Units in the first LSTM layer	{32, 64, 96, 128}	128
Dropout rate after the first LSTM layer	{0.2, 0.3, 0.4, 0.5}	0.3
Units in the second LSTM layer	{32, 64, 96, 128}	32
Dropout rate after the second LSTM layer	{0.2, 0.3, 0.4, 0.5}	0.2
Units in the GRU layer	{32, 64, 96, 128}	128
Dropout rate after the GRU layer	{0.2, 0.3, 0.4, 0.5}	0.3
Units in the dense layer	{32, 64, 96, 128}	64
Dropout rate after the dense layer	{0.2, 0.3, 0.4, 0.5}	0.3
Optimization objective	Validation accuracy	Highest
Hyperparameter search algorithm	Hyperband	-
Initial learning rate	0.001 (Adam default)	Fixed
Learning rate schedule	Exponential decay after epoch 10	Applied

UNCORRECTED PROOF

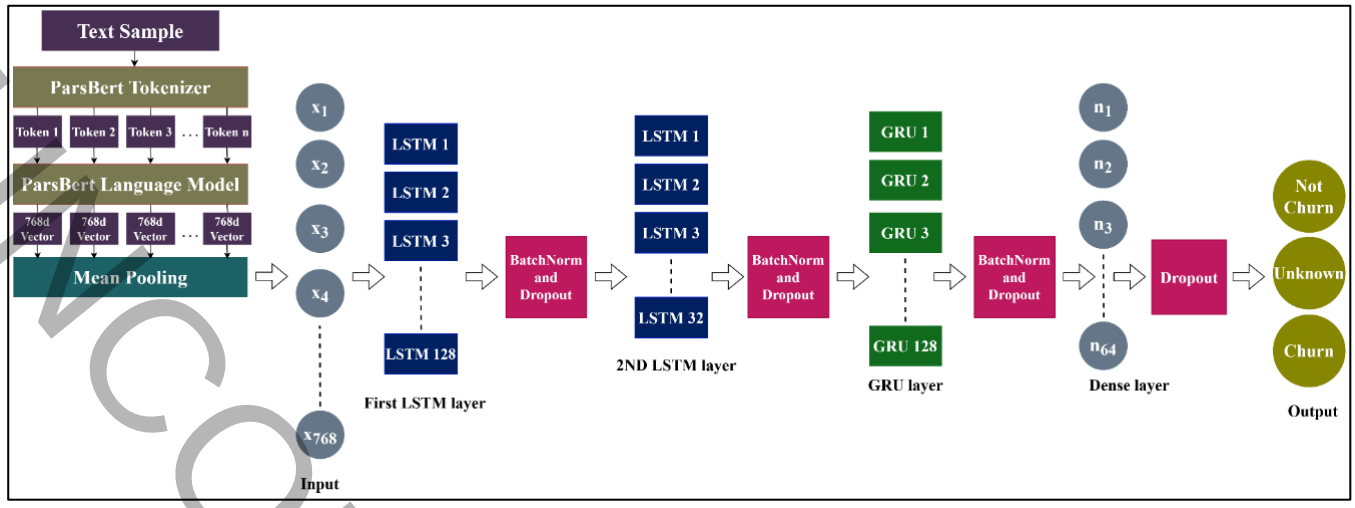


Fig. 2. The architecture of the churn prediction model.

with possible repetition, as each model independently selects from the same unlabeled pool. This stage consists of six main components, each contributing to the process of selecting, labeling, and retraining the model with new data.

• Uncertainty-based Review Selection

In this study, uncertainty-based review selection was performed using the least confidence sampling strategy, where reviews with the lowest prediction confidence were chosen for expert labeling. In this approach, each of the six trained models independently selects 100 reviews from the pool of 9,250 unlabeled Digikala reviews for expert annotation, based on the model's lowest prediction confidence. Specifically, for each model, the class probabilities of all 9,250 unlabeled reviews are predicted. The confidence score for each review is defined as the highest predicted class probability among the three classes ("*Not-Churn*", "*Unknown*", and "*Churn*"). For example, if a model predicts that a particular review belongs to the "*Churn*" class with 60% probability and assigns the remaining 40% across the other two classes, the confidence score for that review is considered to be 60%. Subsequently, the reviews are sorted in ascending order according to their confidence scores. To avoid selecting samples predominantly from a single class, a minimum quota was enforced for each class based on the model's predicted labels. Specifically, at least 20 low-confidence reviews were selected from each of the three classes ("Not-Churn", "Unknown", and "Churn"), while the remaining samples were chosen from the overall pool of the most uncertain reviews regardless of class. This class-aware least-confidence strategy ensures that informative samples from all classes are represented during expert annotation while preserving the advantages of uncertainty-based sampling. As a result, 100 reviews were selected for manual labeling by each model.

This class-aware uncertainty sampling strategy aims to focus expert labeling efforts on the most ambiguous cases while maintaining adequate representation of all classes, thereby providing more informative samples for model improvement. It is worth noting that, for all six models, the confidence scores of the selected samples were below 52%.

• Generating Explanations

After identifying the 100 lowest-confidence reviews for each of the six models, interpretive explanations were generated to assist experts during manual labeling. The main objective of incorporating explainable artificial intelligence (XAI) was to provide insights into the model's uncertainty and facilitate more informed annotation decisions. To this end, the LIME (Local Interpretable Model-agnostic Explanations) algorithm was employed using the *lime* Python library.

LIME explains individual predictions by generating locally perturbed versions of each review and observing how the output probabilities of the original deep learning model change. In this study, no separate classifier was trained for explainability purposes. Instead, the trained LSTM and GRU models based on ParsBERT embeddings were directly used through the prediction interface provided to the LIME library. For textual perturbation, LIME randomly removes words from the original review and queries the corresponding deep model to obtain prediction probabilities for the perturbed samples. A locally interpretable surrogate model, implemented as Ridge regression, is then fitted to these perturbed instances to identify the words that contribute most to the prediction. Therefore, the Ridge regression model serves only as a local approximation and does not replace the original deep learning classifier.

For sequence preparation, each review was first transformed into contextual embeddings using ParsBERT. Since the lengths of the reviews varied, the resulting embedding sequences were padded with zero vectors to a fixed maximum length before being fed into the LSTM and GRU models. During the explainability process, the perturbed

UNCORRECTED PROOF

texts generated by LIME were passed through the same preprocessing pipeline, including tokenization, embedding extraction, and sequence padding, ensuring consistency between model training and explanation generation. In this study, for each trained model, 100 low-confidence reviews underwent this explainability process, and LIME highlighted the top five influential words for each sample to assist experts during manual annotation.

• Expert Labeling

In this step, domain experts review each low-confidence customer opinion together with its corresponding LIME-generated explanation. For every review, LIME identifies and displays the five most influential words contributing to the model's prediction. Based on these highlighted words and the original review text, experts assign the final label to each instance, selecting one of the three predefined categories: Churn, Uncertain, or Non-churn. This labeling process completes one iteration of the active learning cycle before the newly labeled data are added to the training set for model retraining.

• Evaluation of Explanation Usefulness

In this stage, we examine the effectiveness of LIME in supporting experts in labeling customer reviews. For both the randomized and balanced splits, experts were asked to rate the explanations generated for each review on a scale from 1 to 5, where 1 indicated strong dissatisfaction, and 5 indicated high satisfaction with the provided explanations. Specifically, explanations were generated for 587 reviews in the randomized splitting and for 558 reviews in the balanced splitting that were selected through active learning. Experts then assigned scores to the explanations produced for these reviews.

Afterwards, the average scores and the 95% confidence interval (CI) were calculated for each expert in both the randomized and balanced settings to assess the level of satisfaction with the generated explanations. Additionally, experts were asked to label the reviews twice, once before seeing the explanations and once after, in order to determine the difference in inter-annotator agreement between the two stages, and to evaluate whether the explanations contributed to improving agreement among experts.

• Re-training the Churn Prediction Model

After the expert labeling phase, the newly annotated reviews are incorporated into the training dataset of each model. Since the experimental design follows a 6-fold cross-validation setup, this retraining process is performed six times, once for each model corresponding to a specific training–test split. In each iteration, the model is retrained using the expanded training set that includes both the original labeled data and the newly labeled instances obtained through active learning. This step enables the model to update its parameters and refine its decision boundaries based on the newly integrated, informative samples before proceeding to the evaluation phase.

• Evaluation on the Test Fold

Following the retraining step, the updated model is evaluated using the test fold defined in the 6-fold cross-validation setup. In each iteration, the test fold remains unseen during training to ensure an unbiased performance assessment. The model's predictive capability is measured using standard evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics are calculated separately for each fold, and the overall performance is averaged across all six iterations to obtain a robust estimate of the model's generalization ability.

• Semi-Supervised Learning

Following the application of active learning and explainable artificial intelligence (XAI) for labeling 100 samples per model across the six trained models, the next step involves labeling the remaining unlabeled samples using semi-supervised learning. To this end, three different approaches are employed, and the results of each are compared systematically. The first two approaches correspond to self-training and label propagation, which are widely used and well-established methods in the semi-supervised learning literature. The third approach involves a hybrid method that combines the strengths of both self-training and label propagation. This stage consists of three main components:

• Automatic Labeling and Dataset Expansion

In this phase, the model automatically assigns pseudo-labels to the unlabeled reviews based on its prediction confidence. To reduce the risk of introducing noisy labels, only predictions with a maximum class probability greater than or equal to 0.8 were accepted as pseudo-labels. This confidence threshold was determined empirically using a validation set, where different threshold values were evaluated, and 0.8 was selected as the best trade-off between pseudo-label quality and model performance. The labeling process was performed iteratively over 12 stages, allowing the gradual expansion of the labeled dataset. After each stage, the newly labeled samples were incorporated into the training set for model refinement. In the randomized splitting scenario, where the class distribution across folds was not balanced, a total of 8,663 unlabeled reviews were available. During stages 1 through 11, 722 reviews were pseudo-labeled per stage, and 721 reviews were labeled in the final stage. In the balanced splitting scenario, where each fold contained an equal number of samples per class, there were 8,692 unlabeled reviews. In this case, 725 reviews were pseudo-labeled in each of the first 11 stages, and 717 reviews in the final stage, ensuring that all reviews were ultimately labeled. After each stage, the newly labeled samples were incorporated into the training set, gradually expanding the dataset used for model refinement.

• Re-training the Churn Prediction Model

Following each labeling stage, the churn prediction model corresponding to each fold was retrained using the expanded dataset that now included both the original expert-labeled reviews and the newly pseudo-labeled samples.

UNCORRECTED PROOF

This retraining process was repeated at the end of each of the 12 stages, allowing the model to progressively adjust its parameters as new data were added.

• Evaluation on the Test Fold

After each retraining iteration, the model was evaluated using the same test fold that remained unseen during training. The evaluation metrics, including accuracy, precision, recall, and F1-score, were computed after every stage to monitor the model's performance and the progression of the labeling process. This stage-wise evaluation enabled the assessment of how the incremental addition of pseudo-labeled samples influenced the model's predictive capability across the 12 semi-supervised learning iterations.

The pseudo-code of the proposed hybrid semi-supervised learning approach is presented below to clarify the sequence of operations and improve the reproducibility of the method:

Algorithm: Hybrid Semi-Supervised Learning
Input: Labeled dataset L and unlabeled dataset U

Repeat until U is empty:
1. Train the classifier using L.
2. Select unlabeled samples with prediction confidence ≥ 0.8.
3. Randomly divide the selected samples into two equal subsets.
4. Assign pseudo-labels to the first subset using self-training.
5. Assign labels to the second subset using label propagation.
6. Merge the newly labeled samples.
7. Add them to L and remove them from U.
8. Retrain the classifier.

Output: Expanded labeled dataset.

Following the implementation of the described methods, the final phase of this study involves the analysis of the obtained results. By results analysis, we refer to the evaluation metrics recorded at each stage of the research and the extent to which these metrics improve compared to the previous stage. In addition, the outcomes of all three semi-supervised learning methods will be evaluated and compared under both random and balanced splitting. These comparative analyses and the corresponding results will be thoroughly presented and discussed in the following section.

4- Results

In this section, the results of the implementations described in the previous section are analyzed. Overall, this section begins with identifying the influential and important words for each class in the initial labeled dataset. Subsequently, the values of the evaluation metrics after each implementation step are examined separately, indicating how these metrics have changed following each stage. Moreover, the effectiveness of active learning, explainable artificial intelligence, and semi-supervised learning in improving these evaluation metrics is assessed.

4- 1- Frequent Words by Class Label

After preparing the initial labeled dataset consisting of 2,004 samples, where the number of samples in each of the three classes ("*Not-Churn*", "*Unknown*", and "*Churn*") is equal and amounts to 668, it was decided to identify and display the most frequent words in each class. The goal was to determine which words appeared more frequently and had greater influence in each of the three classes. This can be observed in Figures 3 to 5. It is worth noting that for each class, the top 100 most frequent



Fig. 3. Most Frequent Terms in the "Not-Churn" Class.

UNCORRECTED PROOF

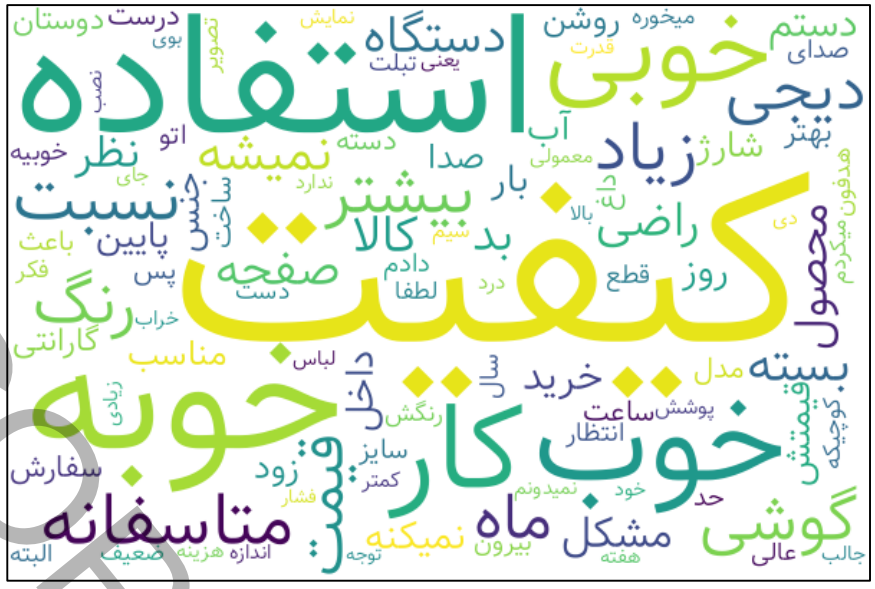


Fig. 4. Most Frequent Terms in the "Unknown" Class.



Fig. 5. Most Frequent Terms in the "Churn Class.

words were selected, and common Persian stop words such as "که" ,"با" ,"از", and similar terms were excluded from the analysis.

As illustrated in the three figures above, the *Not-Churn* class is predominantly characterized by positively connoted words such as «عالی»، «خوب»، «راضی» and similar terms, which occur more frequently than other words. In contrast, the *unknown* class contains a more diverse set of words, including both positive and negative terms. This is attributed to the fact that users often mention both favorable and unfavorable aspects of a product in their reviews, without clearly indicating whether they are genuinely satisfied with the brand or manufacturer. Furthermore, statements that only express personal dissatisfaction without explicitly mentioning the intention not to repurchase from the brand or company in the future are classified under the *Unknown* category. This distinction aligns with the study's focus on predicting customer churn, where only reviews explicitly indicating future non-purchase behavior are labeled as *churn*.

In the *Churn* class, the most frequent words carry

UNCORRECTED PROOF

a negative sentiment, such as «ضعیف», «نخرید», «نمی‌خرم», «پشیمونم» and others, reflecting customer dissatisfaction and their intention to discontinue purchasing, i.e., churn. It is noteworthy that words like «واقعا», «کیفیت» and «خریدم» appear frequently across all three categories, which is expected since such neutral terms do not inherently convey sentiment but rather derive their meaning from surrounding words. Additionally, some words with both positive and negative connotations, such as «بد», «خوب» and «متاسفانه», are observed in all classes. This phenomenon arises due to the common usage of these words.

4- 2- Generated Explanations with LIME

Figures 6 to 8 illustrate examples of the explanations generated by LIME for samples selected through the active learning process.

Figures 6 to 8 illustrate examples of explanations generated using LIME for the samples selected through active learning. These explanations highlight the words within each instance that, according to LIME, played the most significant role in the model's classification decision. For example, in Figure 8, the words «ضعیفه»، «نخرید»، «برند» and «دانید» are identified as having the strongest positive contribution toward the predicted class of *Churn*, whereas the word «بسیار» appears as a negatively contributing term (i.e., against *Churn*). However, due to the stronger influence of churn-supporting words, the final classification remains unchanged.

In general, LIME operates by identifying both the features that support the model's assigned label and those that oppose it. These are visually represented side by side to help the user understand the rationale behind the model's decision. For instance, in Figure 3, LIME identifies four words as supporting the predicted class of *Not-churn* and one word as opposing it. The explanation interface includes a prefix "Not" when referencing the opposing class label. Therefore, in this case, the word contributing against the not-churn label is shown under "Not not-churn", which effectively results in a double negative (i.e., "Not not-churn") being displayed.

An important point worth emphasizing in this context, clearly observable in Figures 6 to 8, is the discrepancy between the model's original confidence score and the class probabilities presented by LIME in its explanations.

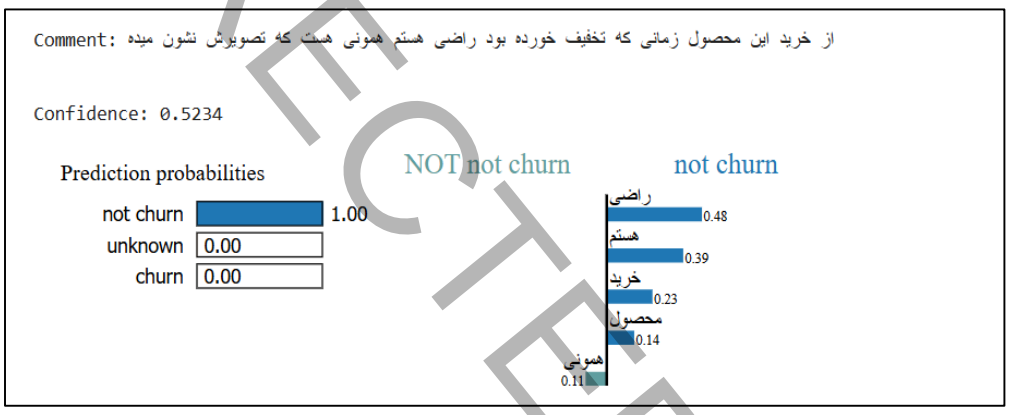


Fig. 6. Example of generating explanations for the "not-churn" class.

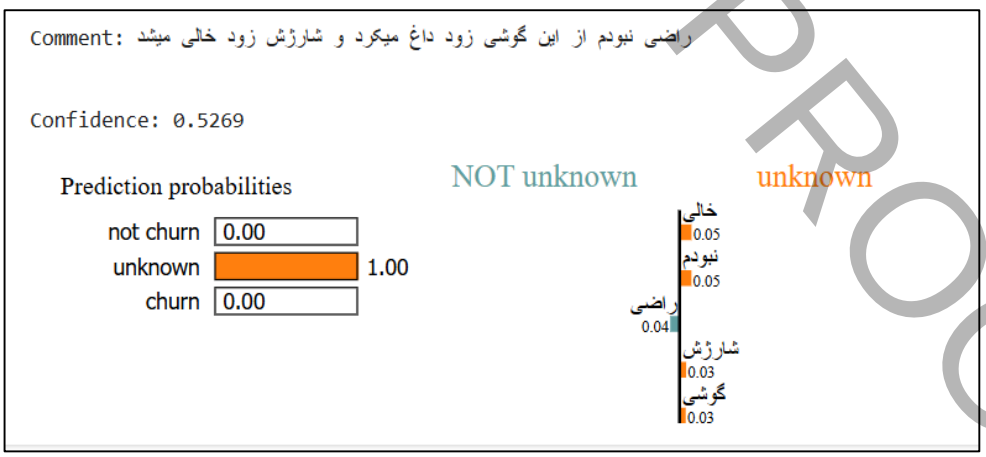


Fig. 7. Example of generating explanations for the "unknown" class.

UNCORRECTED PROOF

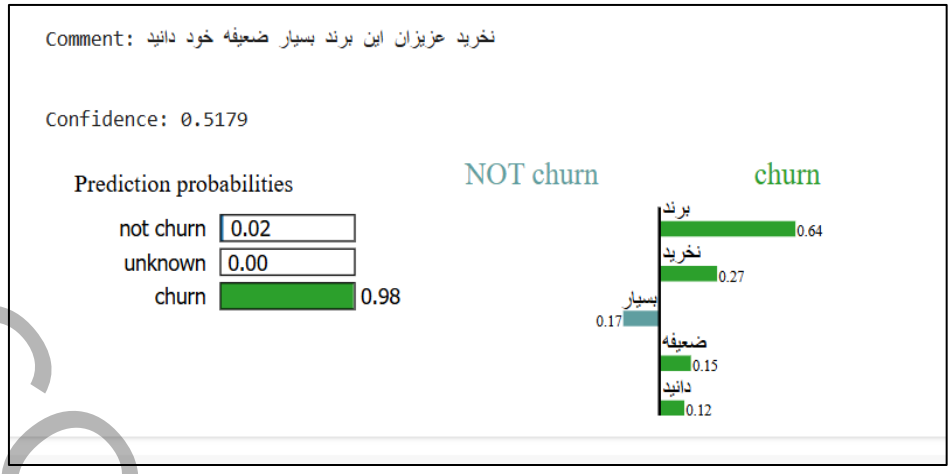


Fig. 8. Example of generating explanations for the "churn" class.

For instance, in Figure 8, the deep learning model assigns a confidence of 51% to the predicted class for a particular sample. However, LIME assigns a significantly higher class probability, 98% for the *Churn* class, in its explanation for the same instance. At first glance, this difference might seem contradictory, as one might expect the two values to be consistent. However, they originate from fundamentally different processes. The model's confidence score is derived directly from the deep learning model's softmax output when predicting the class label. In contrast, LIME generates class probabilities based on a locally interpretable surrogate model trained around perturbed versions of the original input.

As explained earlier, LIME perturbs the original input (e.g., by removing or altering words in the text) to generate a set of neighboring samples, then observes the original model's behavior on these perturbed instances. It uses this behavior to train a simple, interpretable model (typically a linear model) to approximate the decision boundary *locally* around the input. The probabilities produced by LIME represent the surrogate model's interpretation of how strongly each class is supported in the perturbed neighborhood, not the original model's global confidence. Therefore, LIME's class probabilities should not be interpreted as equivalent to the deep model's own confidence scores. If the perturbations happen to emphasize features strongly associated with a particular class (such as *churn*), the surrogate model may heavily favor that class, even if the original model was more uncertain overall. This divergence is natural and expected, given the methodological differences between the two systems.

4- 2- 1- Evaluation of Explanations

As discussed in Section 3.5, the usefulness of the explanations generated by LIME was evaluated. Table 6 reports the results for the randomized and balanced splitting.

According to Table 6, the evaluation results demonstrate that all three experts provided consistently high ratings to the explanations generated by LIME, across both the randomized and balanced splitting[1]. This indicates that the experts found the explanations informative and satisfactory, regardless of the evaluation condition. Notably, the agreement criterion among the experts increased after reviewing the explanations in both settings. This improvement suggests that the explanations not only supported individual expert understanding but also facilitated greater consensus across the group. In other words, the LIME-generated explanations contributed to aligning expert judgments and improving inter-rater reliability, which highlights their potential value in enhancing both interpretability and shared understanding. Collectively, these results provide a direct response to **RQ2**, namely whether the generated explanations were beneficial and effective for experts. The evidence clearly supports this conclusion, highlighting the value of the explanations in improving both individual satisfaction and inter-expert agreement.

4- 3- Churn Prediction Evaluation

In this section, the evaluation metrics are calculated after each phase of the study. As previously mentioned, Fleiss' kappa was employed to assess the inter-expert agreement on the seed dataset containing 2,004 customer reviews, yielding a value of 0.86. This result demonstrates a high level of agreement among the three experts involved in this study. The implementation was conducted under two different conditions for the initial labeled dataset. In the randomized setting, samples were distributed unevenly across the six created folds, resulting in varying class proportions within each fold. Conversely, in the balanced setting, samples were evenly distributed so that the number of samples per class in each fold was approximately equal. The procedure for both

1 The reported agreement improvement corresponds to Fleiss' Kappa values measured before and after providing LIME explanations. The initial Fleiss' Kappa values before using XAI were 0.83 and 0.85 for the randomized and balanced splitting modes, respectively, which increased to 0.91 and 0.92 after incorporating LIME explanations.

UNCORRECTED PROOF

Table 6. Results of Expert Ratings for Evaluation of LIME Usefulness.

Splitting	Experts	Mean Ratings	95% CI	Agreement Improvement
Randomized Splitting	Expert 1	4.165	[4.090, 4.240]	0.83 to 0.91
	Expert 2	4.135	[4.058, 4.211]	
	Expert 3	3.901	[3.823, 3.980]	
Balanced Splitting	Expert 1	4.007	[3.954, 4.061]	0.85 to 0.92
	Expert 2	4.079	[4.026, 4.132]	
	Expert 3	3.812	[3.757, 3.867]	

Table 7. Evaluation Metrics Results After Model Training in the Randomized Splitting.

Model	Accuracy	Precision	Recall	F1-Score
Model 1	82.33%	82.23%	82.33%	82.26%
Model 2	81.73%	82.15%	81.73%	81.85%
Model 3	81.13%	80.98%	81.13%	80.94%
Model 4	84.43%	84.91%	84.43%	84.55%
Model 5	82.93%	83.29%	82.93%	82.81%
Model 6	82.33%	82.42%	82.33%	82.37%
Average	82.48%	82.66%	82.48%	82.46%
Standard deviation	1.14%	1.33%	1.14%	1.2%

settings begins with evaluating the performance of the six churn prediction models trained on the initial labeled dataset of 2004 samples. Each of the six folds serves as a test set for one of the six models, such that each model is tested on a distinct fold. Subsequently, the models are re-evaluated after augmenting their training sets with 100 newly labeled samples obtained through active learning combined with explainable AI. This evaluation uses the same fixed test folds, enabling a clear assessment of the impact of the additional labeled data on model performance. Following this, the semi-supervised learning phase is assessed. Three semi-supervised approaches, self-training, label propagation, and a hybrid method combining the two, are applied, and their results are computed separately for each method. Initially, the results for the randomized scenario are presented and discussed. After completing this, the outcomes for the balanced scenario are reported. Finally, a comparative analysis between these two setups is provided, highlighting the most effective semi-supervised learning strategy.

4- 3- 1- Randomized Splitting

In this section, the results for the randomized splitting, where the number of samples per class is not balanced across folds, are calculated and discussed. First, the evaluation metrics obtained after training and implementing the churn prediction model on the initial labeled dataset will be presented. As previously mentioned, the initial labeled dataset was divided into six folds, with each fold serving once as the test set while the remaining folds were used for training. Table 7 summarizes the evaluation metrics for each of the six models.

As shown in Table 7, the average values of the evaluation metrics exceed 82%, which is considered an acceptable baseline performance. The objective moving forward is to examine how these metrics change after incorporating additional labeled samples obtained through active learning and semi-supervised learning into the training set of each of the six models. The subsequent steps will analyze whether these additions lead to performance improvements.

Active Learning

In the next step, evaluation metrics are calculated for each model after incorporating the labeled samples obtained through active learning and explainable artificial intelligence (XAI). As previously mentioned, 100 additional labeled samples generated through active learning and XAI are added to the training set of each of the six models to assess the extent

UNCORRECTED PROOF

Table 8. Evaluation Metrics Results After Applying Active Learning in the Randomized Splitting.

Model	Accuracy	Precision	Recall	F1-Score
Model 1	83.53%	83.43%	83.53%	84.44%
Model 2	84.43%	84.65%	84.43%	84.51%
Model 3	84.43%	84.44%	84.43%	84.38%
Model 4	85.32%	85.69%	85.32%	85.43%
Model 5	85.32%	85.54%	85.32%	85.23%
Model 6	85.32%	85.55%	85.32%	85.41%
Average	84.72%	84.88%	84.72%	84.73%
Standard deviation	0.73%	0.88%	0.73%	0.78%

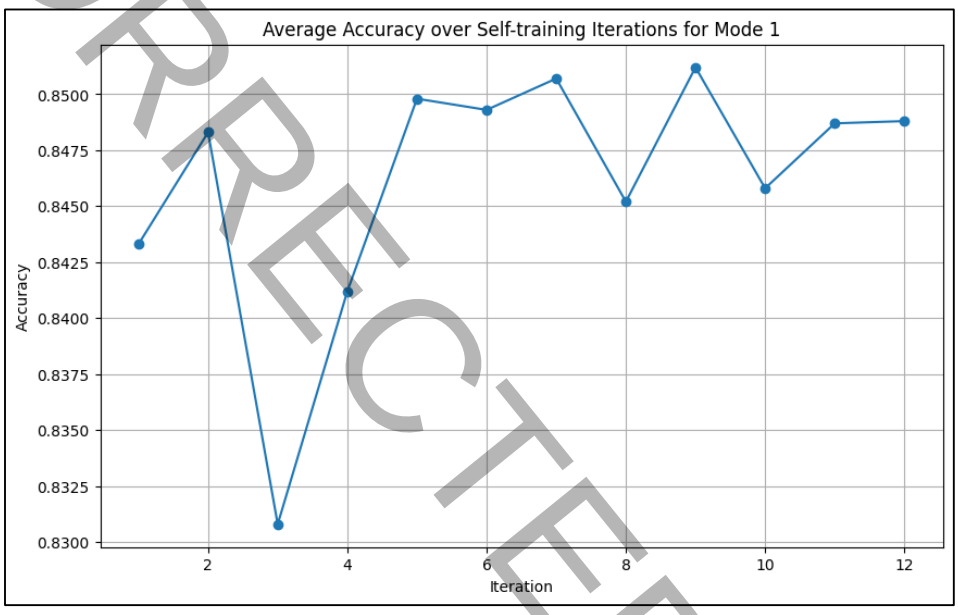


Fig. 9. Average Accuracy Across the Six Models using the Self-training Method under Randomized Splitting.

of improvement in the evaluation metrics. It is important to note that the test set for each of the six models remains unchanged, consisting of the original 334 samples, while the training set size increases from 1,670 to 1,770 samples. Table 8 presents the evaluation metrics for each of the six models after the inclusion of the active learning and XAI-generated samples in their respective training datasets.

As presented in Table 8, after adding 100 labeled samples to the training set of each of the six models using active learning and explainable AI, the average evaluation metrics improved by more than 2.2%. This demonstrates that the application of active learning and explainable AI has led to an enhancement in model performance. Considering the relatively small number of added samples, this level of improvement can be regarded as both reasonable and expected.

In the next step, the evaluation metrics are calculated after adding the labeled samples, obtained through semi-supervised learning, to the training dataset of each of the six models. As previously mentioned, three semi-supervised learning approaches are employed: self-training, label propagation, and a hybrid method that combines the first two. The labeling process is carried out in 12 sequential stages for each model, with 722 samples labeled in each stage (except for the final stage, which includes 721 samples). After each stage, the evaluation metrics are computed to observe how the performance evolves. The results corresponding to each of the three approaches are presented and analyzed in the following sections.

Self Training

In this section, the evaluation results for the self-training method are presented. Figure 9 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

UNCORRECTED PROOF

Table 9. Best Evaluation Metric Results in Self-Training Method in the Randomized Splitting.

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	6, 9	6012, 8268	84.13%	84.19%	84.13%	84.01%
Model 2	5	5380	85.33%	85.59%	85.33%	84.43%
Model 3	11	9712	86.23%	86.18%	86.23%	86.19%
Model 4	7, 8	6824, 7546	86.83%	86.88%	86.83%	86.85%
Model 5	8, 9	7546, 8268	87.13%	87.41%	87.13%	87.02%
Model 6	5	5380	87.43%	87.62%	87.43%	87.42%
Average			86.18%	86.31%	86.18%	86.15%
Standard deviation			1.25%	1.29%	1.25%	1.26%

As illustrated in Figure 9, the accuracy metric exhibits a fluctuating pattern across the 12 self-training stages, reflecting the varying quality of the pseudo-labeled samples incorporated at each iteration. Since some pseudo-labels may introduce noise, temporary increases and decreases in performance are expected. To justify the selection of the optimal stage and avoid choosing a stage with inferior performance, a model selection criterion based on validation accuracy was employed. Specifically, after each stage, the models were evaluated on the validation set, and the stage with the highest average validation accuracy across the six models was selected. According to this criterion, stage 9 achieved the best performance, with an average accuracy of 85.12%. The trends observed for the other three evaluation metrics were similar; therefore, their corresponding plots have been omitted. The same procedure was applied to the subsequent methods as well. Table 9 presents the highest achieved value of each evaluation metric for every individual model.

As shown in Table 9, except for Model 3, the best evaluation metrics for the remaining models were obtained during the intermediate stages. Moreover, as indicated by the average row, the evaluation metrics in the self-training method improved by approximately 1.5% compared to the active learning approach. The column titled *"Number of Samples"* in the table refers to the total number of training samples. For instance, in Model 2, the total number of training samples, including the initial labeled samples, the 100 samples added via active learning, and those added through the self-training method, amounts to 5,380 samples.

Label Propagation

In this section, the evaluation results for the label propagation method are presented. Figure 10 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

As illustrated in Figure 10, the highest average accuracy was achieved at Stage 7, reaching 84.03%. This indicates that among the six models, the average accuracy at Stage 7 attained the optimal value. In this chart, the average accuracy exhibits a downward trend from Stage 7 onward, with a consistent decline at each subsequent stage. This observation suggests that the Label Propagation method, within the context of this study, is not well-suited for labeling large quantities of samples and is only effective up to a certain point. Table 10 presents the highest achieved value of each evaluation metric for every individual model.

As shown in Table 10, and consistent with the previously described pattern, the highest accuracy values for the six models were predominantly achieved during the early stages, with the exception of Model 1, which reached its peak accuracy in a mid-stage. This outcome was not observed in the later stages, indicating that the performance of the Label Propagation method declines as the number of stages increases.

Hybrid Method

After completing the labeling process using the *Self-Training* and *Label Propagation* methods, the *Hybrid* method, combining these two approaches, was employed for labeling the remaining samples. This section presents and discusses the evaluation results obtained from the *Hybrid* method. Figure 11 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

As illustrated in Figure 11, the highest mean accuracy was achieved in the first stage, reaching 84.23%. This indicates that, on average, the six models attained their best accuracy during the initial stage. From the very beginning, however, the mean accuracy exhibited a consistent downward trend, with a gradual decline observed across subsequent stages. This finding suggests that, within the context of this study, the *Hybrid* method is not a suitable approach for large-

UNCORRECTED PROOF

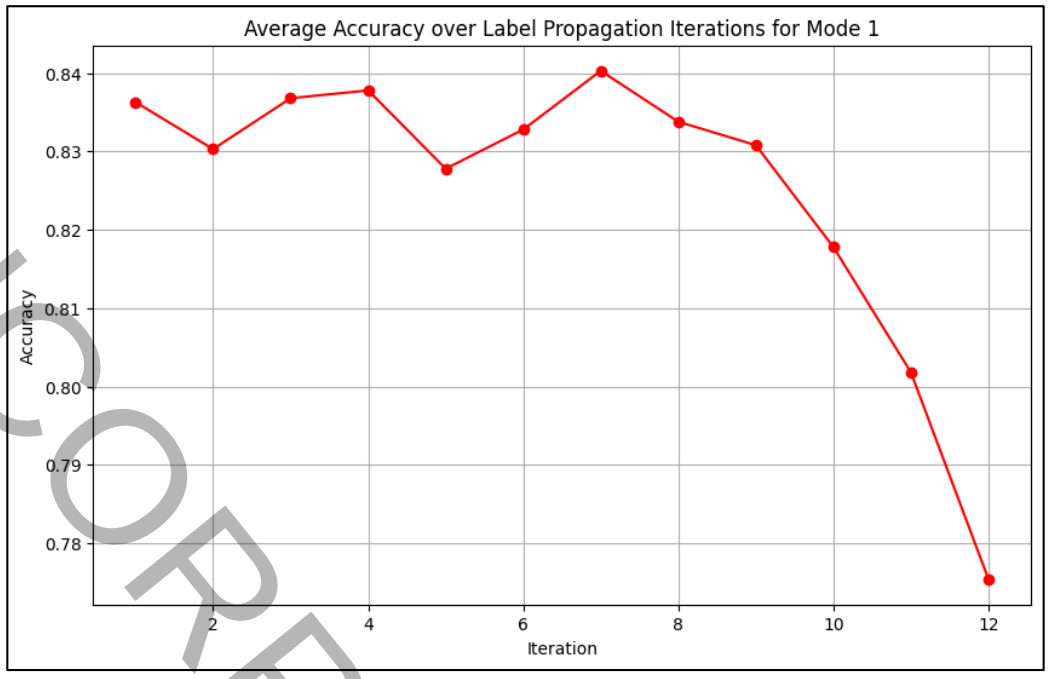


Fig. 10. Average Accuracy Across the Six Models using the Label Propagation Method under Randomized Splitting.

Table 10. Best Evaluation Metric Results in Label Propagation Method in the Randomized Splitting

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	7	6824	83.53%	83.49%	83.53%	83.47%
Model 2	4	4658	84.73%	84.76%	84.73%	84.74%
Model 3	3	3936	84.43%	84.35%	84.43%	84.26%
Model 4	4	4658	86.53%	86.93%	86.53%	86.62%
Model 5	4, 6	4658, 6012	85.33%	85.70%	85.33%	85.22%
Model 6	2	3214	85.93%	86.06%	85.93%	85.98%
Average			85.08%	85.21%	85.08%	85.04%
Standard deviation			1.08%	1.25%	1.08%	1.15%

scale sample labeling and cannot be considered a reliable strategy. Table 11 presents the highest achieved value of each evaluation metric for every individual model.

As shown in Table 11, the highest evaluation metric values for the six models were achieved in either stage 1 or stage 2. This observation indicates that, as the number of stages increased, the metric values declined compared to the early stages. Therefore, the *Hybrid* method demonstrated satisfactory accuracy only at the initial stages of the process.

In the following section, the results for the balanced setting will be presented. After examining the outcomes of this setting, a comparison between the results of the two settings will be conducted, followed by a comprehensive conclusion based on the findings.

4- 3- 2- Balanced Splitting

In this section, the results for the balanced setting, where the number of samples per class is balanced across folds,

UNCORRECTED PROOF

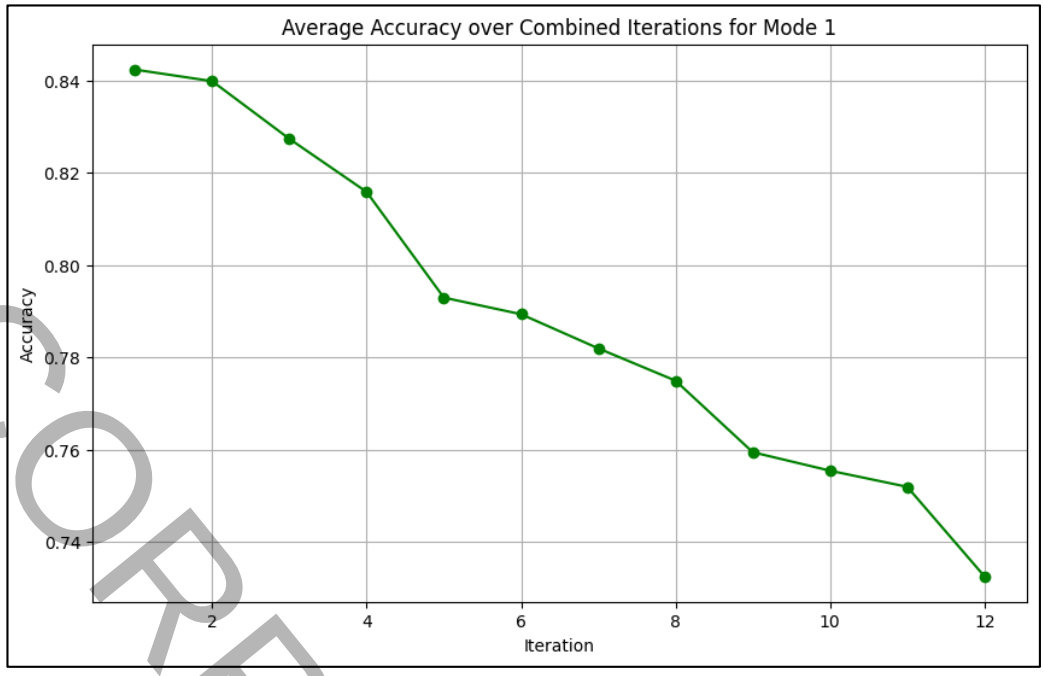


Fig. 11. Average Accuracy Across the Six Models using the Hybrid Method under Randomized Splitting

Table 11. Best Evaluation Metric Results in Hybrid Method in the Randomized Splitting.

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	1	2492	83.53%	83.40%	83.53%	83.44%
Model 2	2	3214	84.43%	84.70%	84.43%	84.54%
Model 3	2	3214	85.03%	85.05%	85.03%	85.03%
Model 4	2	3214	85.33%	85.33%	85.33%	85.41%
Model 5	1	2492	84.73%	84.73%	84.73%	84.66%
Model 6	1	2492	85.93%	86.24%	85.93%	86.03%
Average			84.83%	85.01%	84.83%	84.85%
Standard deviation			0.82%	0.95%	0.82%	0.88%

are calculated and discussed. First, the evaluation metrics obtained after training and implementing the churn prediction model on the initial labeled dataset will be presented. As previously mentioned, the initial labeled dataset was divided into six folds, with each fold serving once as the test set while the remaining folds were used for training. Table 12 summarizes the evaluation metrics for each of the six models.

As shown in Table 12, the average values of the evaluation metrics are very close to 83%, which, similar to the randomization mode, can be considered acceptable as a baseline. The objective moving forward is to examine how these metrics change after incorporating additional labeled samples obtained through active learning and semi-supervised learning into the training set of each of the six models. The subsequent steps will analyze whether these additions lead to performance improvements.

In the next step, evaluation metrics are calculated for each model after incorporating the labeled samples obtained

UNCORRECTED PROOF

Table 12. Evaluation Metrics Results After Model Training in the Balanced Split.

Model	Accuracy	Precision	Recall	F1-Score
Model 1	82.93%	82.92%	82.93%	82.86%
Model 2	85.03%	85.09%	85.03%	85.04%
Model 3	82.93%	82.87%	82.93%	82.78%
Model 4	80.84%	80.73%	80.84%	80.76%
Model 5	81.74%	81.84%	81.74%	81.78%
Model 6	84.43%	84.29%	84.43%	84.29%
Average	82.98%	82.96%	82.98%	82.92%
Standard deviation	1.58%	1.58%	1.58%	1.57%

Table 13. Evaluation Metrics Results After Applying Active Learning in the Balanced Splitting.

Model	Accuracy	Precision	Recall	F1-Score
Model 1	83.53%	83.43%	83.53%	84.44%
Model 2	84.43%	84.65%	84.43%	84.51%
Model 3	84.43%	84.44%	84.43%	84.38%
Model 4	85.32%	85.69%	85.32%	85.43%
Model 5	85.32%	85.54%	85.32%	85.23%
Model 6	85.32%	85.55%	85.32%	85.41%
Average	84.72%	84.88%	84.72%	84.73%
Standard deviation	0.73%	0.88%	0.73%	0.78%

through active learning and explainable artificial intelligence (XAI). As previously mentioned, 100 additional labeled samples generated through active learning and XAI are added to the training set of each of the six models to assess the extent of improvement in the evaluation metrics. It is important to note that the test set for each of the six models remains unchanged, consisting of the original 334 samples, while the training set size increases from 1,670 to 1,770 samples. Table 13 presents the evaluation metrics for each of the six models after the inclusion of the active learning and XAI-generated samples in their respective training datasets.

Active Learning

As shown in Table 13, after adding 100 samples to the training dataset of each of the six models using active learning and explainable artificial intelligence, the average values of the evaluation metrics increased by more than 2%. Similar to the randomization mode, the use of active learning and explainable AI led to improved churn prediction performance. Considering the number of samples added, this level of improvement can be regarded as both reasonable and expected.

In the next step, the evaluation metrics are calculated after adding the labeled samples, obtained through semi-supervised learning, to the training dataset of each of the six models. As previously mentioned, three semi-supervised learning approaches are employed: self-training, label propagation, and a hybrid method that combines the first two. The labeling process is carried out in 12 sequential stages for each model, with 725 samples labeled in each stage (except for the final stage, which includes 717 samples). After each stage, the evaluation metrics are computed to observe how the performance evolves. The results corresponding to each of the three approaches are presented and analyzed in the following sections.

UNCORRECTED PROOF

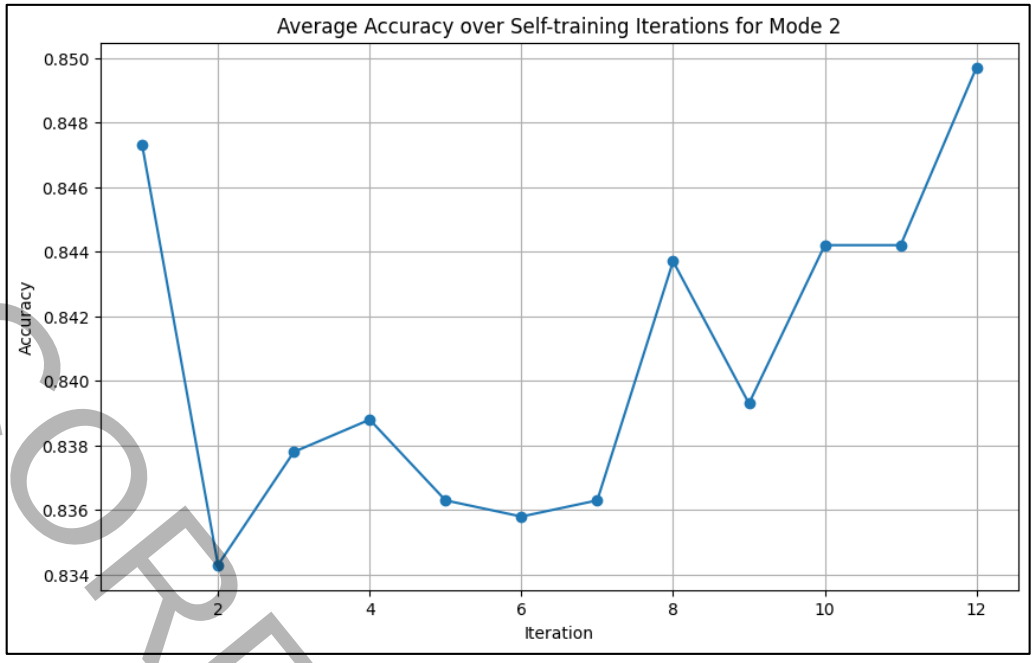


Fig. 12. Average Accuracy Across the Six Models using the Self-training Method under Balanced Splitting.

Table 14. Best Evaluation Metric Results in Self-Training Method in the Balanced Splitting.

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	3	3945	87.13%	87.02%	87.13%	86.98%
Model 2	9	8295	87.13%	87.29%	87.13%	87.05%
Model 3	1	2495	85.63%	85.63%	85.63%	85.58%
Model 4	12	10462	83.23%	83.51%	83.23%	83.29%
Model 5	11, 12	9745, 10462	84.13%	84.46%	84.13%	84.23%
Model 6	1, 10	2495, 9020	87.13%	87.07%	87.13%	86.95%
Average			85.73%	85.83%	85.73%	85.68%
Standard deviation			1.71%	1.57%	1.71%	1.61%

Self Training

In this section, the evaluation results for the self-training method in a balanced setting are presented. Figure 12 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

As illustrated in Figure 12, the trend resembles a sinusoidal pattern, similar to the randomization mode, with no discernible consistent pattern. In this figure, for each of the 12 stages, the average accuracy across the six models was computed. The highest average accuracy was achieved in the final stage, reaching 84.97%, indicating that stage 12 yielded the best average performance among the six models. Table 14 presents the highest achieved value of each evaluation metric for every individual model.

As presented in Table 14, the highest evaluation metrics for the six models were obtained at different stages, which supports the previously noted sinusoidal pattern of accuracy variations across the stages. The optimal accuracy could be achieved at any of the 12 stages; for instance, model 3 reached its highest accuracy in stage 1, whereas model 4 achieved its best performance in the final stage. Furthermore, averaging the best accuracies of all six models revealed an approximate

UNCORRECTED PROOF

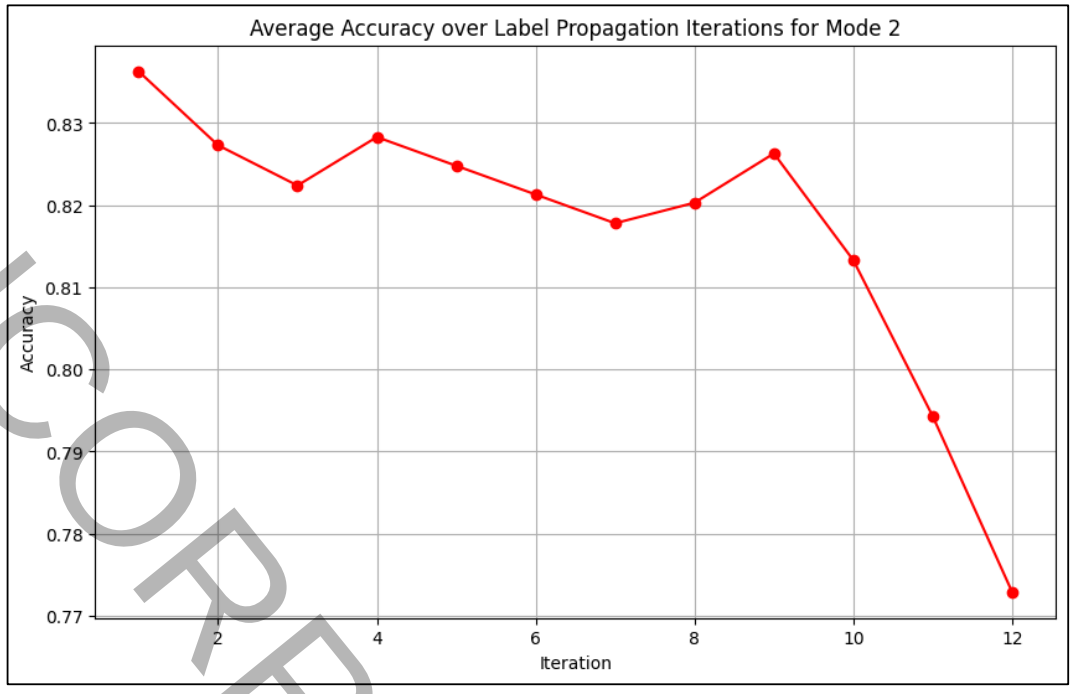


Fig. 13. Average Accuracy Across the Six Models using the Label Propagation Method under Balanced Splitting.

Table 15. Best Evaluation Metric Results in Label Propagation Method in the Balanced Splitting.

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	4	4670	86.83%	86.85%	86.83%	86.79%
Model 2	4	4670	85.63%	85.73%	85.63%	85.66%
Model 3	1	2495	85.93%	86.08%	85.93%	85.98%
Model 4	2	3220	80.84%	80.94%	80.84%	80.83%
Model 5	1	2495	80.84%	80.79%	80.84%	80.72%
Model 6	2	3220	86.83%	86.75%	86.83%	86.78%
Average			84.48%	84.52%	84.48%	84.46%
Standard deviation			2.86%	2.86%	2.86%	2.89%

0.5% improvement compared to the mean obtained after applying active learning.

Label Propagation

In this section, the evaluation results for the label propagation method are presented. Figure 13 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

As illustrated in Figure 13, the highest mean accuracy was achieved in stage 1, reaching 83.63%, indicating that the average accuracy of the six models peaked at the initial stage. From stage 9 onward, the mean accuracy exhibited a downward trend, with a consistent decrease observed at each subsequent stage. Similar to the results obtained under the randomization setting, this pattern suggests that the label propagation method, within the context of this study, is not well-suited for labeling large volumes of data and is only effective up to a certain extent. Table 15 presents the highest achieved value of each evaluation metric for every individual model.

As presented in Table 15 and consistent with the previously described pattern, the highest accuracy values

UNCORRECTED PROOF

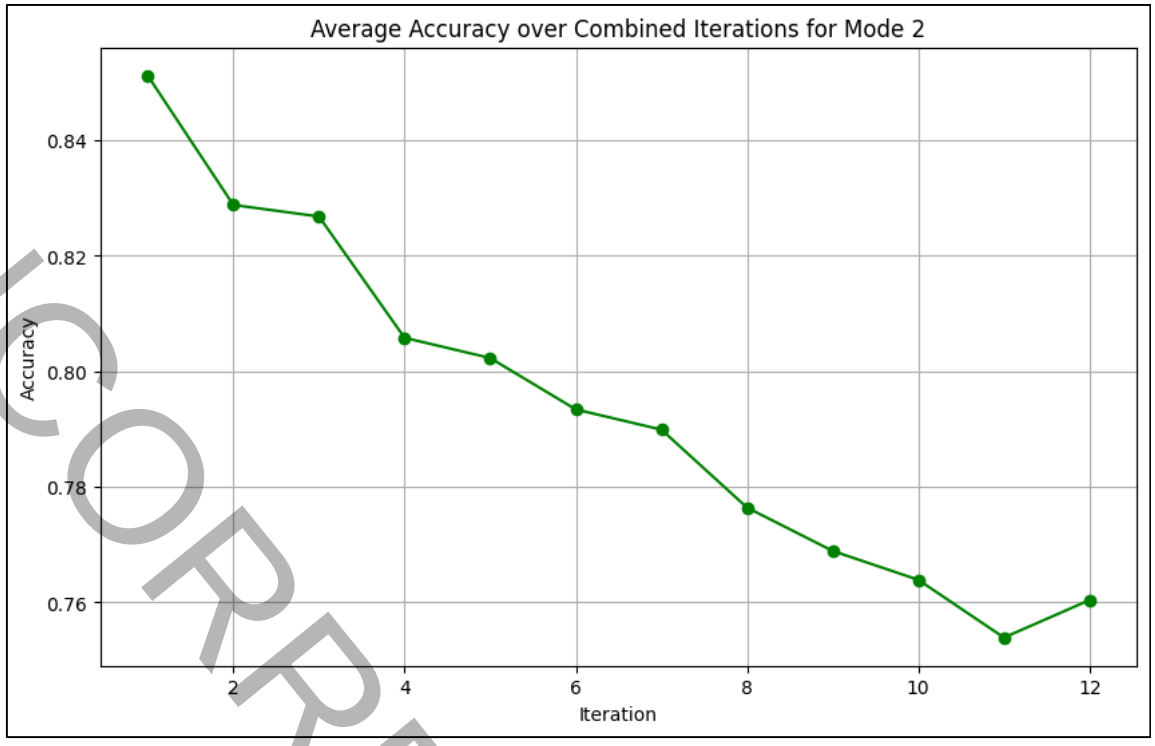


Fig. 14. Average Accuracy Across the Six Models using the Hybrid Method under Balanced Splitting.

for the six models under the balancing setting, similar to the randomization setting, were predominantly achieved in the early stages. This observation indicates that as the number of stages increases, the label propagation method progressively loses its effectiveness. Given that these results align closely with those obtained under the randomization setting, it can be concluded with greater confidence that label propagation is only effective for labeling up to a certain threshold. Beyond this point, it tends to reduce evaluation metrics and ultimately degrade model performance.

Hybrid Method

After completing the labeling process using the *Self-Training* and *Label Propagation* methods, the *Hybrid* method, combining these two approaches, was employed for labeling the remaining samples. This section presents and discusses the evaluation metric results obtained from the *Hybrid* method. Figure 14 shows the average accuracy across all six models at each stage of the semi-supervised learning process.

As illustrated in Figure 14, the highest mean accuracy was obtained at stage 1, with a value of 85.12%. This indicates that, for stage 1, the average accuracy across the six models reached its maximum. Similar to the randomization setting, the mean accuracy began to decline from the very early stages, exhibiting a consistently downward trend. This finding suggests that, within the context of this study, the combined method is not a suitable approach for sample labeling and cannot be considered a reliable strategy. Table 16 presents the highest achieved value of each evaluation metric for every individual model.

As shown in Table 16, the highest values of the evaluation metrics for all six models in the balanced setting were achieved entirely at stage 1. This indicates that, as the number of stages increases, the metric values decrease compared to the initial stages, and that the combined method delivers high accuracy only at the very beginning, specifically in stage 1.

4- 3- 3- Comparison of the Results in the Randomized and Balanced Splitting

Table 17 presents a comparison between the results of the randomized and balanced splitting.

Table 17 shows that the application of active learning and semi-supervised learning improved accuracy in both the randomized and balanced settings, thereby providing affirmative answers to RQ1 and RQ3 regarding the effectiveness of these techniques in enhancing performance. As shown in Table 17, the results obtained under the randomized and balanced settings are largely comparable, with both demonstrating the best performance when employing the self-training method. Although the average of the best accuracies across the six models in the self-training method was slightly higher in the randomized setting (86.18%) than in the balanced setting (85.73%), the difference was relatively small. To investigate whether this difference

UNCORRECTED PROOF

Table 16. Best Evaluation Metric Results in Hybrid Method in the Balanced Splitting.

Model	Stage	# of Samples	Accuracy	Precision	Recall	F1-Score
Model 1	1	2495	85.33%	85.30%	85.33%	85.29%
Model 2	1	2495	87.13%	87.13%	87.13%	87.08%
Model 3	1	2495	85.63%	85.62%	85.63%	85.61%
Model 4	1	2495	82.34%	82.58%	82.34%	82.37%
Model 5	1	2495	83.23%	83.13%	83.23%	83.15%
Model 6	1	2495	87.13%	87.10%	87.13%	87.02%
Average			85.13%	85.14%	85.13%	85.08%
Standard deviation			1.98%	1.93%	1.98%	1.96%

Table 17. Comparison of Results between the Randomized and Balanced Splitting.

Comparison factor	Randomized Splitting	Balanced Splitting
1- Mean initial accuracies of the six models	82.48%	82.98%
2- Mean accuracies of the six models after applying active learning	84.72%	85.07%
3- Mean of the best accuracies of the six models in the self-training method	86.18%	85.73%
4- Mean of the best accuracies of the six models in the label propagation method	85.08%	84.48%
5- Mean of the best accuracies of the six models in the hybrid method	84.83%	85.13%
6- Best mean accuracy of the six models in the self-training method	85.12% Stage 9	84.97% Stage 12
7- Best mean accuracy of the six models in the label propagation method	84.03% Stage 7	83.63% Stage 1
8- Best mean accuracy of the six models in the hybrid method	84.23% Stage 1	85.12% Stage 1
9- Best semi-supervised learning method based on the results	Self-Training	Self-Training

was statistically significant, a paired t-test was conducted using the best accuracy values of the six models under both settings. The resulting p-value was 0.688, indicating that the observed difference was not statistically significant at the 0.05 significance level. In addition, a Wilcoxon signed-rank test yielded a p-value of 0.625, leading to the same conclusion. These findings suggest that the proposed framework exhibits stable performance under both randomized and balanced data partitioning schemes, demonstrating its robustness with respect to different splitting strategies.

Furthermore, considering the best average accuracies of the six models for each method (Items 6–8 in Table 17), the randomization scenario achieved its highest value with self-training in stage 9, while the balancing scenario attained its highest value with the combined method in stage 1. However, in both cases, the improvements relative to the post–active

UNCORRECTED PROOF

learning accuracies were not substantial. It should be noted that, although the combined method achieved the highest average accuracy in stage 1 for the balancing scenario, the corresponding accuracy trend, as illustrated in Figure 14 and discussed earlier, showed a consistent decline from the early stages onward. This suggests that the combined method is unsuitable for large-scale sample labeling, as its effectiveness is limited to the initial stages. Therefore, self-training remains the more reliable and effective approach for this task.

Overall, when comparing the three semi-supervised learning methods applied in this study, self-training, label propagation, and the combined method, the self-training approach emerges as the most effective. As evidenced by the results for both experimental scenarios, the accuracy metric under self-training does not exhibit a notable decline when labeling a large number of samples. Even in later stages, with a substantial volume of labeled data, the accuracy remains at an acceptable level. In contrast, for label propagation, as shown in Figures 10 and 13, the accuracy begins to decrease from the mid-stages onward. For the combined method, as illustrated in Figures 11 and 14, the accuracy trend is consistently downward, indicating that this approach is unsuitable for large-scale labeling tasks. Consequently, the answer to **RQ4**, is clearly self-training.

The superiority of self-training over the other two methods can be attributed to its reliance on the model's own predictions to generate labels, which keeps it closely aligned with the original supervised learning signal. Moreover, it selectively incorporates high-confidence predictions, thereby reducing the likelihood of introducing noise. In contrast, label propagation assigns labels based on sample similarity within a graph structure, which can introduce noise over successive iterations. In this method, small initial errors may propagate and amplify, leading to a gradual degradation in accuracy. The combined method, which merges self-training and label propagation, is even more vulnerable to performance deterioration, as it blends two inherently conflicting strategies. This vulnerability arises because label propagation assumes a globally smooth label distribution, whereas self-training relies on the model's prediction confidence, two principles that may interfere with one another when applied simultaneously.

A key distinction lies in the fact that self-training remains model-driven and adaptive, preserving the original learning signal, whereas label propagation and the combined method are more susceptible to error accumulation and drift from the initial training distribution. Additionally, it is important to note that the Persian language exhibits complex morphology and has relatively fewer high-quality, large-scale linguistic resources compared to many other languages. This limitation can hinder the ability of textual representations, such as word embeddings and their transformation into numerical vectors, to capture semantic similarities with high precision. Consequently, the similarity graph used in label propagation may be noisy or weakly structured, allowing incorrect labels to spread more easily between dissimilar or borderline samples. While pre-trained language models for Persian are steadily improving, they still tend to be less robust than their English counterparts, meaning that word embeddings or similarity scores may be noisier or less semantically reliable. As a result, graph edges may inadvertently connect samples from different classes, facilitating the propagation of erroneous labels.

Overall, the results indicate that applying active learning in combination with semi-supervised learning (specifically, the self-training method) for labeling Persian-language customer reviews aimed at churn prediction can be an effective approach. Throughout the labeling process, the evaluation metrics not only avoided degradation but also showed improvements, suggesting that this method can serve as a reliable alternative to reduce high costs and dependence on human annotators. Furthermore, no substantial differences were observed between the randomized and balanced settings, as their results were largely similar.

4- 4- Discussion

This research investigated the effectiveness of three semi-supervised learning approaches, self-training, label propagation, and a hybrid method, for large-scale labeling of Persian-language textual data in the context of customer churn prediction. The experiments were conducted under two sampling strategies: randomization and balancing, and the process was augmented with active learning and explainable artificial intelligence (XAI) to guide sample selection and labeling.

Across both experimental settings, the self-training approach consistently achieved the highest evaluation metrics and demonstrated greater stability over iterative labeling cycles. Unlike label propagation and the hybrid method, self-training did not exhibit a notable decline in accuracy as more samples were labeled and integrated into the training set. In fact, in both random and balanced scenarios, accuracy improvements were observed even in later stages, indicating that the approach is resilient to the noise that typically accumulates in semi-supervised labeling processes.

The superior performance of self-training can be attributed to its model-driven, adaptive nature. By iteratively selecting high-confidence predictions from the model itself, self-training effectively preserves the original supervised learning signal and limits the propagation of early-stage errors. This confidence-based filtering mechanism is particularly well-suited to domains where data quality and representation can vary significantly, as it mitigates the risks of mislabeling that often undermine other semi-supervised methods.

In contrast, label propagation applies a graph-based strategy, diffusing labels across samples according to pairwise similarity scores. While theoretically appealing, this approach is more vulnerable to error amplification, especially in high-iteration contexts, because mislabels can spread across the graph to dissimilar or borderline samples. This vulnerability is exacerbated in Persian text classification due to linguistic and resource-related factors. Persian's rich morphological structure, combined with the scarcity of large-scale, high-quality linguistic resources, can limit the accuracy

UNCORRECTED PROOF

of word embeddings or other vector-based representations. Consequently, similarity graphs for Persian text may contain noisy or weakly structured edges, leading to incorrect label diffusion. Although Persian pre-trained language models have improved in recent years, they remain less robust than their English counterparts, increasing the likelihood that similarity-based labeling introduces semantic drift over time.

The hybrid method, designed to combine self-training and label propagation, was expected to leverage the strengths of both approaches. However, the results reveal that it instead inherits their weaknesses. The method combines conflicting operational assumptions, global label smoothing from propagation and local, confidence-based selection from self-training, resulting in decreased model performance, particularly when labeling large volumes of data. In both random and balanced sampling scenarios, the hybrid method's performance declined steadily after the initial iterations, confirming its unsuitability for sustained large-scale labeling in this domain.

A distinctive feature of this study was the integration of active learning with XAI into the labeling process. Active learning enabled the prioritization of informative and uncertain samples, ensuring that each new batch of labeled data maximized the model's learning potential. Meanwhile, XAI techniques, specifically model interpretability tools such as LIME, were employed to make the labeling process transparent and justifiable. This allowed annotators (human-in-the-loop) to validate model predictions based on feature-level explanations before adding them to the labeled dataset. The combination of model confidence from self-training, informativeness from active learning, and interpretability from XAI provided a multi-layered quality assurance mechanism, significantly reducing the risk of label noise in the final training set.

When comparing randomization and balancing strategies, both yielded broadly similar patterns: self-training was the best-performing method, followed by label propagation, with the hybrid approach consistently underperforming. The main quantitative difference was that in the randomization setting, self-training achieved slightly higher gains in accuracy (a 3.7% improvement over initial averages) compared to balancing (2.75%). This suggests that while balancing can be useful for ensuring class representation, it does not inherently enhance the performance of semi-supervised methods in this task, and in some cases, the diversity introduced by randomization may be more beneficial.

From a practical perspective, the findings strongly support the adoption of self-training combined with active learning and XAI for labeling large-scale Persian textual datasets. This integrated approach avoids the performance degradation seen in other methods, maintains transparency for human validators, and minimizes annotation costs by reducing reliance on manual labeling. Furthermore, the method's robustness across both sampling strategies suggests it can be reliably applied in operational churn prediction pipelines without the need for extensive pre-tuning of data selection strategies.

From a theoretical standpoint, the results highlight several implications for semi-supervised learning in morphologically rich, low-resource languages. First, similarity-based methods like label propagation are particularly sensitive to the quality of linguistic embeddings; without high-quality semantic representations, graph-based diffusion risks propagating structural noise rather than useful information. Second, hybridizing methods with incompatible inductive biases may not yield additive benefits and can instead compound weaknesses. Third, incorporating explainability into semi-supervised pipelines not only improves trust and interpretability but also plays a crucial role in controlling error propagation, a consideration that has received little attention in prior literature.

5- Conclusion

In this study, Persian customer reviews were annotated using a combination of active learning and semi-supervised learning to facilitate customer churn prediction. As noted in the introduction, most previous research on churn prediction has focused on structured data, particularly in the telecommunications domain, with limited work addressing unstructured textual data, especially in the Persian language. Therefore, this research aimed to address the gap by developing a churn prediction framework specifically for Persian textual data.

A major challenge encountered was the absence of publicly available labeled datasets for Persian churn prediction. To address this, we adopted a multi-stage annotation strategy integrating active learning, explainable artificial intelligence (XAI), and semi-supervised learning. Initially, a labeled dataset of 2,004 Persian text samples was constructed with the expertise of three human annotators, categorized into three classes: *not-churn*, *unknown*, and *churn*. A baseline churn prediction model was trained on this dataset. Subsequently, an unlabeled dataset collected from Digikala was processed using active learning augmented with XAI to select informative samples for manual annotation by domain experts. The remaining unlabeled samples were then annotated using semi-supervised learning, employing three approaches: self-training, label propagation, and a hybrid method combining the two.

Evaluation results demonstrated that combining active learning with semi-supervised learning, particularly the self-training approach, proved effective for annotating large-scale Persian textual data without significant degradation in evaluation metrics. In fact, in many cases, this approach led to modest improvements in accuracy compared to the post–active learning stage. This outcome suggests that the proposed framework can substantially reduce annotation costs and reliance on human labor, while still producing high-quality labeled data at scale. Importantly, the findings also revealed that churn prediction is fundamentally distinct from customer satisfaction prediction, and the two tasks should not be conflated in modeling approaches. The novelty of this research lies primarily in its methodology: the integration of active learning with XAI for sample selection, followed

UNCORRECTED PROOF

by semi-supervised labeling of Persian textual data for churn prediction, a strategy not previously reported for this language or task. Additionally, the creation of the first labeled dataset for Persian churn prediction constitutes a significant contribution, filling a key resource gap for future research.

Overall, the proposed methodology, leveraging active learning, XAI, and semi-supervised learning, provides an effective and innovative approach for labeling Persian text data for churn prediction. The results indicate that this approach is reliable, scalable, and cost-efficient, making it a practical solution for real-world applications where labeled data are scarce and expensive to obtain.

Despite the promising results achieved in this study, several limitations should be acknowledged. In particular, only a single active learning strategy, namely least confidence (uncertainty) sampling, was employed for selecting informative samples. Although this strategy demonstrated satisfactory performance, the present study did not investigate whether alternative query strategies could provide superior results. Therefore, the comparative effectiveness of other active learning approaches, such as margin-based sampling and entropy-based sampling, remains an open question. Addressing this limitation could provide a deeper understanding of the impact of sample selection strategies on the quality of labeled Persian textual data and the subsequent performance of churn prediction models.

For future research in this direction, several avenues can be explored to further improve and extend the proposed framework. A particularly important direction is the systematic comparison of different active learning strategies for annotating Persian reviews. In this study, the least confidence sampling method was adopted and yielded satisfactory results. However, future studies should investigate and compare alternative strategies, such as margin-based sampling and entropy-based sampling, to determine whether they can further improve annotation quality and model performance. Such comparisons would provide a more comprehensive understanding of the role of query strategies in active learning-based annotation frameworks.

Another possible direction is to enhance the performance of the *label propagation* method for annotation. Our results indicated that, unlike self-training, label propagation was unable to deliver desirable and robust results for large-scale datasets. Future research could examine potential modifications or hybridization techniques that would enable label propagation to effectively handle high-volume Persian text annotation.

Additionally, exploring imbalanced learning scenarios could provide valuable insights. In this study, the initial dataset was perfectly balanced, with an equal number of samples in each of the three classes. Future experiments could utilize imbalanced datasets and apply methods such as minority class oversampling or majority class undersampling to assess the robustness of the proposed framework under class imbalance conditions.

Finally, the methodology developed in this study could be applied to domains beyond online store reviews. For example, the framework could be adapted for analyzing patient reviews of doctors or client reviews of lawyers, to predict client attrition in those professional contexts. Such extensions would not only test the generalizability of the approach but also broaden its applicability to other real-world text-based churn prediction problems.

Author contributions: Soroush Bajoghli:

Conceptualization, Software, Validation, Investigation, Data Curation, Writing- Original Draft. **Fakhroddin Noorbehbahani:** Conceptualization, Supervision, Methodology, Writing - Original Draft, Writing- Review and Editing.

Funding:

This research received no external funding.

Data Availability

The raw customer review corpus used in this study was obtained from the Digikala dataset [34]. The processed, labeled dataset used for experiments, including the expert-verified seed set and the expanded labels produced via our active learning and semi-supervised procedures, along with annotation guidelines and example LIME explanations, are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest: The authors declare no conflict of interest.

References

[1] S. M. Keaveney and M. Parthasarathy, "Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors," Journal of the Academy of Marketing Science, vol. 29, no. 4, pp. 374-390, 2001.

[2] F. Reichheld, "Zero Defections: Quality Comes to Services," Harvard Business Review, 1990.

[3] R. T. Rust and A. J. Zahorik, "Customer satisfaction, customer retention, and market share," Journal of Retailing, vol. 69, no. 2, pp. 193-215, 1993.

[4] N. N. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," Knowledge-Based Systems, vol. 212, p. 106586, 2021.

[5] C.-F. Tsai and Y.-H. Lu, "Data mining techniques in customer churn prediction," Recent Patents on Computer Science, vol. 3, no. 1, pp. 28-32, 2010.

[6] Tomanek, K., & Hahn, U. (2009, August). Semi-supervised active learning for sequence labeling. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 1039-1047).

[7] Nguyen, V. L., Shaker, M. H., & Hüllermeier, E. (2022). How to measure uncertainty in uncertainty sampling for

UNCORRECTED PROOF

active learning. Machine Learning, 111(1), 89-122.

[8] Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC medical informatics and decision making, 25(1), 110.

[9] S. A. A. Asli, B. Sabeti, Z. Majdabadi, P. Golazizian, R. Fahmi, and O. Momenzadeh, "Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian," in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 2855-2861.

[10] Y. Huang et al., "Telco churn prediction with big data," in Proceedings of the 2015 ACM SIGMOD international conference on management of data, 2015, pp. 607-618.

[11] K. Coussement and D. Van den Poel, "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers," Expert Systems with Applications, vol. 36, no. 3, pp. 6127-6134, 2009.

[12] H. Amiri and H. Daume III, "Target-dependent churn classification in microblogs," in Proceedings of the AAAI Conference on Artificial Intelligence, 2015, vol. 29, no. 1.

[13] H. Amiri and H. Daume III, "Short text representation for detecting churn in microblogs," in Proceedings of the AAAI Conference on Artificial Intelligence, 2016, vol. 30, no. 1.

[14] M. Gridach, H. Haddad, and H. Mulki, "Churn identification in microblogs using convolutional neural networks with structured logical knowledge," in Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017, pp. 21-30.

[15] A. Giannakopoulos, M. Coriou, A. Hossmann, M. Baeriswyl, and C. Musat, "Resilient combination of complementary CNN and RNN features for text classification through attention and ensembling," in 2019 6th Swiss Conference on Data Science (SDS), 2019: IEEE, pp. 57-62.

[16] J. Zhong and W. Li, "Detecting Customer Churn Signals for Telecommunication Industry Through Analyzing Phone Call Transcripts with Recurrent Neural Networks," in MLDM (1), 2019, pp. 93-103.

[17] C. Abbet et al., "Churn intent detection in multilingual chatbot conversations and social media," arXiv preprint arXiv:1808.08432, 2018.

[18] Z. Rajabi and M. Valavi, "A survey on sentiment analysis in Persian: a comprehensive system perspective covering challenges and advances in resources and methods," Cognitive Computation, vol. 13, no. 4, pp. 882-902, 2021.

[19] L. Taherkhani, A. Daneshvar, H. Amoozad Khalili, and M. R. Sanaei, "Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm," Advances in Civil Engineering, vol. 2023, no. 1, p. 6029121, 2023.

[20] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in 22nd International Conference on Computational Linguistics, Coling 2008, 2008, pp. 1137-1144.

[21] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 917-926.

[22] S. Li, S. Ju, G. Zhou, and X. Lin, "Active learning for imbalanced sentiment classification," in Proceedings of the 2012 Joint conference on empirical methods in natural language processing and computational natural language learning, 2012, pp. 139-148.

[23] Y. Zhang, M. Lease, and B. Wallace, "Active discriminative text representation learning," in Proceedings of the AAAI conference on artificial intelligence, 2017, vol. 31, no. 1.

[24] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using SVM for text classification," International Journal of Automation and Computing, vol. 15, pp. 290-298, 2018.

[25] M. Ghayoomi and J. Kuhn, "Sampling methods in Active Learning for treebanking," in Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT'12), 2013, pp. 49-60.

[26] J. Chen, J. Feng, X. Sun, and Y. Liu, "Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts," Symmetry, vol. 12, no. 1, p. 8, 2019.

[27] Y. Li and J. Ye, "Learning adversarial networks for semi-supervised text classification via policy gradient," in Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 1715-1723.

[28] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake news detection," in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 568-569.

[29] H. Linmei, T. Yang, C. Shi, H. Ji, and X. Li, "Heterogeneous graph attention networks for semi-supervised short text classification," in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 4821-4830.

[30] T. Kumar, J. Park, M. S. Ali, A. S. Uddin, J. H. Ko, and S.-H. Bae, "Binary-classifiers-enabled filters for semi-supervised learning," IEEE Access, vol. 9, pp. 167663-

UNCORRECTED PROOF

167673, 2021.

[31] Z. Jahanbakhsh-Nagadeh, M.-R. Feizi-Derakhshi and A. Sharifi, "A semi-supervised model for Persian rumor verification based on content information," Multimedia Tools and Applications, vol. 80, no. 28, pp. 35267-35295, 2021.

[32] M. Mahmoodvand and M. Hourali, "Semi-supervised approach for Persian word sense disambiguation," in 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), 2017: IEEE, pp. 104-110.

[33] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," Artificial Intelligence Review, vol. 56, no. 9, pp. 9401-9469, 2023.

[34] https://camelcase.ir/wp-content/uploads/digikala_dataset[www.camelcase.ir].zip

[35] Noorbehbahani, F., Bajoghli, S., & Esfahani, H. H. (2023, May). Customer Loyalty Prediction of E-marketplaces Via Review Analysis. In 2023 9th International Conference on Web Research (ICWR) (pp. 311-316). IEEE.

HOW TO CITE THIS ARTICLE
S. Bajoghli, F. Noorbehbahani, Labeling Text Reviews for Customer Churn Prediction via Active Learning, Semi-Supervised Learning, and Explainable AI, AUT J. Model. Simul., 58(2) (2026) 217-250.

DOI: 10.22060/miscj.2026.25707.5487

UNCORRECTED PROOF