BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

An answerer recommendation approach on Stack Overflow using Historical Posts with User Profile Data

Muazul Islam
200042132
Tawfiq-E-Elahi
200042147
Ahmed Mahfuz Anan
200042173

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

An answerer recommendation approach on Stack Overflow using Historical Posts with User Profile Data

Muazul Islam
200042132
Tawfiq-E-Elahi
200042147
Ahmed Mahfuz Anan
200042173

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Muazul Islam**, **Tawfiq-E-Elahi**, and **Ahmed Mahfuz Anan** under the supervision of **Shohel Ahmed**, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Shohel Ahmed
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 30, 2025

Muazul Islam
Student ID: 200042132
Date: September 30, 2025

Tawfiq-E-Elahi
Student ID: 200042147
Date: September 30, 2025

Ahmed Mahfuz Anan
Student ID: 200042173
Date: September 30, 2025

Dedicated to our supervisor for continuous guidance, mentorship, and encouragement during the pursuit of this work.

Contents

List of Figures

List of Abbreviations

CQA	Community Question Answering
NMF	Non-negative Matrix Factorization,
TEM	Topic Expertise Model
TTEA	Temporal Topic Expertise Activity
SMOTE	Synthetic Minority Over-sampling Technique
LightGCN	Lightweight Graph Convolutional Networks
CNN	Convolutional Neural Networks
MRR	Mean Reversed Rank
NDCG	Normalized Discounted Cumulative Gain

Acknowledgement

I would like to express my deepest gratitude to Mr. Shohel Ahmed, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology (IUT), for his insightful guidance, unwavering direction, and continuous encouragement throughout the process of this thesis. His experience, constructive criticism, and useful suggestions have played a pivotal role in the course of this research work.

I am also wholeheartedly thankful to the Department of Computer Science and Engineering, Islamic University of Technology, for providing me with facilities, academic environment, and assistance that greatly helped me perform my research work.

Abstract

The exponential growth of community-driven knowledge platforms such as StackOverflow has created a pressing need for effective recommendation systems that can personalize content to users expertise and interests. This thesis presents a hybrid recommendation approach that integrates semantic topic embeddings, temporal activeness features, and reputation-based contribution metrics into a unified two-tower neural recommendation framework.

To address the challenge of high-dimensional and composite tag vocabularies, a preprocessing strategy was developed to adapt pretrained topic embeddings to over 10,000 StackOverflow tags, including hyphenated and dot-separated forms. User representations were constructed by aggregating embeddings of interacted tags, augmented with features derived from activity patterns and reputation dynamics. Negative sampling was employed using cosine similarity thresholds to differentiate relevant from irrelevant tags, while candidate sampling ensured balanced evaluation across users.

Experimental results across multiple configurations demonstrate that topic embeddings provide a strong semantic baseline, but recommendation quality does not improve with the addition of activeness and contribution features. The contribution of this paper is an empirical insight that semantic topic signals dominate, and adding activeness and reputational features naively reduces performance. The finding is contradictory to the assumptions of other papers, such as IEA [22], on the fact that using different types of features improves the result.

Chapter 1

Introduction

The rise of digital technologies has significantly transformed the way knowledge is created, shared, and accessed. Community Question Answering (CQA) platforms have emerged as dynamic spaces where users collaborate by asking questions, providing answers, and engaging in discussions. Unlike static knowledge repositories, CQA platforms evolve continuously through active user participation, fostering real-time problem-solving and collective learning.

Stack Overflow [19], established in 2008, has become the foremost CQA platform for programmers and developers worldwide. Hosting millions of technical posts, it serves as a crucial knowledge hub where users build reputations based on the quality of their contributions. Despite its success, Stack Overflow faces persistent challenges, such as a considerable number of unanswered questions and inefficiencies in connecting questions with the right expert users. Maintaining effective knowledge exchange and high user engagement requires addressing these issues.

Tagging plays an important role in organizing content on Stack Overflow, helping users find relevant posts[20]. However, user-assigned tags are often inaccurate or inconsistent, particularly when added by less experienced members. This reduces the discoverability of questions and contributes to the growing backlog of unanswered posts.

Recommendation systems, widely adopted across domains like e-commerce and entertainment, offer promising solutions by personalizing content delivery based on user behavior and preferences. In CQA platforms, intelligent recommendation systems can enhance knowledge sharing by suggesting newly posted or unanswered questions to users best suited to answer them. Such systems not only improve response rates but also strengthen community engagement.

This research draws on the principles of recommendation systems to address some of the existing challenges on Stack Overflow. By improving the matching between questions and expert users, it aims to support a more efficient, dynamic, and responsive knowledge-sharing environment.

1.1 Motivations and Scope

Question-and-answer (Q&A) websites like Stack Overflow are vital to the construction of software today and technical learning. These websites rely considerably on community contribution to generate and sift content, with millions of users contributing answers to a wide range of technical questions. But the effectiveness of such websites depends upon their ability to refer questions to the most appropriate users with the appropriate knowledge and interest. In practice, most questions remain unanswered or are provided less-than-perfect responsesnot because there aren't enough knowledgeable users out there, but because the system can't find and recommend the appropriate contributors.

While existing recommendation systems have come a long way using state-of-the-art machine learning models such as neural networks and large language models (LLMs), they overlook user-level signals that are not related to question or answer content. This effort draws inspiration from filling this gap using a user-centric recommendation approach. The focus of this study, therefore, is to create a methodology that rigorously integrates user profile informationsuch as topic knowledge, contribution history, contribution patterns, and platform usageto build a more effective and context-aware answerer recommendation system.

1.2 Problem Statement

Despite advances in technology for recommendation systems, Stack Overflow continues to have a large volume of unanswered or poorly answered questions. One of the root reasons for this issue is the mismatch of questions with users who lack sufficient expertise or interest in the specific topic. Current models primarily focus on content-based features or advanced learning techniques, without considering rich user profile information that can provide more insight into a user's capability to contribute meaningfully.

Therefore, the primary problem addressed by this research is: How do we improve an-

swerer recommendation in Stack Overflow by incorporating rich user profile features including expertise and activity-based behavior, alongside existing content-based approaches?

1.3 Research Challenges

Building a proficient recommendation system for experts at Stack Overflow has several subtle challenges. One of the key challenges is the Stack Overflow data dump, in which many attribute values are null or missing, making it difficult for one to accurately discover the user's profile with respect to all their past contributions. Recent user updates like changes in badges and reputation can be retrieved through the Stack Overflow API, but the volume possible is restricted due to strict API rate limits and the ever-increasing user population. This necessitates partial updates, which make it impossible to build fully up-to-date user profiles.

In addition, another challenge is the inherent inconsistency in using a combination of static data from the dump and dynamic data from the API that makes it impossible to replicate real time behavior of a user, thus affecting the fidelity of the evaluation of the recommendation framework. The challenges posed by the nature of data are modeled user activeness, which is ironically difficult because the mode of engagement widely varies over the course of time. Finally, extracting accurate semantic information from question texts in Stack Overflow is made difficult because informal phrasing-for example, ambiguity or non-standard detail-appears to be common among the queries posted. All these points emphasize the complexity in building an expert recommendation system that will be scalable and robust at Stack Overflow.

1.4 Contribution

This research makes the following contributions to the question-answer recommendation field:

- A User-Centered Recommendation Framework: We propose a new framework that departs from content-based models to a user profiling approach that captures topic knowledge, activity, activity pattern, and platform contribution measures.
- Integration of User Profile Attributes: Unlike with the numerous earlier attempts that fail to consider user profile data, our solution integrates statistics such as answer acceptance record, badge hierarchical levels, reputation trend,

and topic-based contribution, which together offer a closer estimate of user capability.

- A Practical Scoring Model: We design an explainable scoring system that combines a few metrics to rank users so that more accurate and reliable recommendations are made possible without relying on computationally expensive deep learning models.
- Identification of Gaps in Current Literature: From a thorough literature review, we identify the lack of integration of user profiles in current approaches and create a need for an equilibrated approach that considers both content and user-specific information.

We make this research work towards building a better, scalable, and interpretable system for answerer recommendation, with possibilities of extension to other community platforms.

1.5 Organization

This thesis is organized into several chapters to ensure a logical and coherent flow of ideas.

Chapter 1, *Introduction*, outlines the background, motivations, problem statement, research challenges, and contributions of the study.

Chapter 2, *Related Works*, reviews existing literature on expert recommendation systems, topic modeling, and user profiling, identifying gaps this research addresses.

Chapter 3, *Proposed Methodology*, describes the system design, methodological components, and their integration, supported by architectural diagrams.

Chapter 4, *Dataset Discussion*, discusses the datasets used, highlights data limitations, and explains how these challenges were handled.

Chapter 5, *Conclusion*, summarizes the findings, discusses the implications, acknowledges limitations, and suggests directions for future work.

Together, these chapters provide a structured journey from problem identification to solution development and reflection.

Chapter 2

Related Works

2.1 Semantic Clustering Model

Huang et al. [7] proposed this approach which combines semantic embeddings (Word2Vec) and collaborative filtering (Non-negative Matrix Factorization, NMF) to suggest CQA specialists. It preprocesses the posts in advance, maps text into TF-weighted Word2Vec vectors (new semantic-aware representation), and groups them into implicit knowledge areas using k-means (clustering technique) to prevent noisy user-labeled tags[7]. For recommendation, it covers content-based similarity (through domain clusters) and NMF-based (traditional collaborative filtering) latent user-answer interaction scores to counteract data sparsity. Early studies focused on link analysis and user behavior (Jurczyk and Agichtein [8];), while graph-based browsing models addressed language dependence and sparsity issues (Chiang et al. [1]). Other approaches explored question difficulty (Hanrahan et al. [5]), voting behaviors (Zhang et al. [28]), and topic modeling techniques like LDA and STM for user profiling (Riahi et al. [15]). Methods relying on tags (Guo et al. [4]; Dong et al. [2]) often suffered from inaccuracy due to user-generated noise. The paper's novelty is in the hybrid model that combines semantic post analysis and voting behavior of users to maximize the accuracy of recommendation over tag-only or matrix factorization-only models[7]. But limitations are reliance on Stack Overflow voting to identify quality, oversimplification in spaces employing k-means clustering, and vocabulary loss through excessive word filtering (e.g., 1.3M to 5k words), all potentially at the cost of context-specific distinctions[7]. Finally, overly sparse user-answer interactions are a problem for NMF, which acknowledges scalability issues with sparse data.

2.2 Topic-Activity Hybrid Model

Earlier answerer-recommendation methods in CQA (e.g. Yang et al. [25]'s TEM) attempted to infer the topical interests and expertise of the users through latent topic models. Meng et al. [12] generalized this by formulating temporal activity models (TTEA) and an activeness variant (TTEA-ACT) that combine answer and trendbased activity. In contrast, the IEA method calculates for every answer candidate a combined topical interest, expertise, and activeness score[22]. In practice, IEA first makes an educated guess about a user's topic-interest and expertise vectors from his/her past questions and answers (with a TEM-like model) and then calculates an activeness measure that incorporates both answering and commenting activity . The overall score is then calculated by simply multiplying these three components[22]. Comment-based activeness being included is IEA's biggest innovation, and it performs recommendation more accurately than TEM, TTEA or TTEA-ACT. Wang et al. [22] mention that the technique merely multiplies these components without further weighting, and it has been evaluated on StackOverflow data only. Also naive multiplicative scoring exclude complex interactions, vote score is used as ground truth where ideal answerers are assumed.

2.3 Behavioral Classification Model

Roy and Singh [17] focus on identifying future experts at an early stage of their community participation. Instead of ranking established high-reputation users in prior research that required extensive user history, such as ExpertiseRank (Zhang et al. [28]), graph-based ranking (Jurczyk and Agichtein [8]), and hybrid authority reputation models (Yang et al. [25]), their early expert prediction model aims to spot promising experts newcomers who have begun contributing high-quality content using only the first month of user activity[17]. The proposed methodology extracts features from a users initial questions, answers, and feedback (votes) in that initial period and trains a predictive model to forecast whether the user will go on to become a top expert in the community. Empirical evaluation on CQA data (drawn from Stack Exchange archives) shows that this approach can accurately distinguish future experts from ordinary users even with limited observation data[17]. In fact, Roy and Singh [17] report high predictive performance, suggesting that early contribution patterns such as answering frequency, answer acceptance rate, and community appreciation are strong indicators of long-term expertise development. This is a significant methodological shift from knowledge embedding (Yuan et al. [27]), and time-evolving

expertise models (Van Dijk et al. [21]). One benefit of early prediction is enabling platforms to engage rising talents (through incentives or question routing) before unanswered questions pile up[17]. However, the approach assumes that short-term behavior is reliably indicative of future potential; it may miss late-blooming experts or falsely identify users who show early promise but later drop off[17]. The studys positive results, nonetheless, highlight that incorporating temporal dynamics and user growth trajectories can complement existing expert recommendation systems, moving beyond static expertise rankings to a more predictive, lifecycle-oriented view of community experts.

2.4 Comprehensive Review Model

Yuan et al. [27] provide a comprehensive survey of expert finding techniques in community question answering, comparing approaches and identifying trends. They categorize existing solutions into four groups: matrix factorization-based models(Hu et al. [6], Salakhutdinov et al. [18], Liang and De Rijke [11]), gradient boosting decision tree (ensemble) models, deep learning models(Ying et al. [26], Wei et al. [23]), and ranking models. An interesting insight from their survey is that, on a standard benchmark (Toutiaos ByteCup competition data), classic MF-based models outperformed the other categories on average, highlighting the strong performance of latent factor methods in this domain. However, the authors note that top-performing systems often combine multiple model types for example, ensemble methods that blend factorization with neural or tree-based predictors yielded further performance boosts in the ByteCup challenge[27]. The review also discusses how different models fare on various matching scenarios (pure text matching vs. graph-based or multimodal contexts), providing guidance on model selection for practitioners[27]. Moreover, the paper emphasizes emerging considerations unique to CQA, such as user willingness to answer, answer quality, and dynamic behavior, which traditional expert finding methods did not fully address. The survey concludes that while substantial progress has been made e.g., incorporating social graphs and deep contextual language models the field is moving towards hybrid models and fairness-aware, real-time expert recommendation[27]. It identifies open challenges like data sparsity (many new questions and users), and suggests that future research should explore richer user profiling (including temporal activity and incentives) to further improve expert finding in online communities.

2.5 Embedding-Based Semantic Disambiguation Model

Efstathiou et al. [3] present a domain-specific word embedding model aimed at improving semantic understanding in software engineering Q&A contexts. They trained a Word2Vec model on 15 GB of Stack Overflow posts to capture the nuanced meaning of technical terms and jargon in this domain [3]. The resulting vector representations (dubbed SO Word2Vec) can effectively disambiguate polysemous words (e.g., Java as a programming language vs. coffee) by leveraging the software-related contexts present in the data. The authors illustrate that these embeddings encode fine-grained semantic relationships (for example, clustering of library names or error codes), which implies their potential transferability to various information retrieval and mining tasks in software engineering[3]. A notable benefit of the SO-specific embeddings is the boost in relevance when matching technical content, compared to general-purpose embeddings[13] which often miss subtle domain meanings[3]. One limitation is that the model is trained on posts up to 2018 and may require updates as technology and vocabulary evolve. Additionally, these embeddings serve as an improved input representation that can be integrated into broader CQA expert finding frameworks for better semantic matching.

2.6 Domain-Aware Embedding Model

Huang et al. [7] proposes an expert recommendation framework that moves beyond manual tags by learning knowledge domain embeddings from CQA content. Their approach first trains distributed word embeddings on a large archive of Stack Overflow posts, then clusters these embeddings to uncover latent knowledge domains essentially topics or subfields emerging from how terms co-occur in technical discussions [7]. High-quality Q&A pairs are identified and assigned to these latent domains, and users expertise is quantified by their contributions within each domain (e.g. number of accepted answers or upvotes in that domain). To mitigate sparse participation data, they apply matrix factorization to the user-domain activity matrix, inferring missing associations and strengthening the expertise signals. When a new question arises, it is mapped into the same semantic domain space, and the system recommends users with top rankings in the relevant domain. Experiments on the Stack Overflow dataset demonstrated that this method, which the authors extend from an earlier conference version, outperforms baseline models that rely on explicit tags or basic topic models[7]. In particular, it surpassed prior topic modeling approaches (e.g., Riahi et al. [15]'s LDA-based model and Dong et al. [2]'s SSRM) in precision@N and nDCG, in-

dicating more accurate and high-quality expert rankings [7]. The models strength lies in automatically discovering topic structure from data and incorporating content semantics, which improved recommendation stability across varying query loads. A potential limitation is the reliance on unsupervised clustering to define expertise domains the chosen number of clusters and their coherence can impact performance. Nevertheless, this embedding-based approach offered a novel way to capture experts knowledge areas and showed robust gains over state-of-the-art methods on multiple large-scale CQA datasets.

2.7 Graph-Semantic Fusion Model

Wu et al. [24] introduces LG-ERMG, a multi-granularity hybrid method that combines Lightweight Graph Convolutional Networks (LightGCN) with multi-granularity semantic encoding to better support expert recommendation for Community Question Answering (CQA). Classical methods (e.g., BM25[16], Doc2Vec[9]) are based on keyword matching or topic models, whereas more recent neural methods (e.g., CNTN[14], NeRank[10]) employ CNNs or heterogeneous network embeddings but ignore graph-structured relationships. LG-ERMG proposes two innovations. The first one is LightGCN, which reduces regular GCNs by eliminating nonlinear activations and self-loops to consider only graph topology to enhance efficiency and avoid overfitting. And another is a multi-granularity encoder that calculates semantic similarity at word, question, and expert-levels with Transformers and attention mechanisms. This combination captures both structural expert relationships as well as fine-grained semantic matches. The model attains state-of-the-art performance, beating baselines by 515% on NDCG@10 and MRR, showing that integrating graph-based expert interactions with hierarchical semantic analysis greatly improves recommendation accuracy. The model ignores temporal dynamics (e.g., changing expert interests). It also has cold-start issues for new questions/users. It requires careful hyperparameter tuning (e.g., embedding dimension).

2.8 Comparative analysis

The diverse expert recommendation models in CQA platforms adopt fundamentally different strategies. CQARank uses a probabilistic topic model (Topic Expertise Model) to learn latent topic-expertise profiles from question text and user activity, then applies a topic-sensitive PageRank-like algorithm over the question-answer network to rank users by expertise [25]. This joint modeling of topical content and link struc-

ture enables identification of users who not only share topical interests with a new question but have proven authority through community endorsement. Experiments showed CQARank significantly outperformed pure topic models by leveraging both text and network signals [25]. However, CQARanks authors noted that user interests and expertise evolve over time a dimension their model did not capture [25]. The Topic-Activity Hybrid Model (IEA) represents a next step, explicitly integrating a users current activeness with their topical expertise. IEA combines three components long-term topical interest, historical expertise, and short-term activity level to recommend answerers for new questions [22]. By analyzing not just past questions and answers but also comment behavior, IEA captures the intuition that an expert must be both knowledgeable and presently active in the community [22]. In practice, this hybrid modeling of content and recent behavior led to notable performance gains. Wang et al. report that IEA outperforms earlier models focused only on topics or only on activity, achieving higher ranking accuracy and correlation with ground truth experts than methods like TEM or TTEA-ACT [22]. Still, IEAs notion of activeness addresses availability in the very short term and relies on historical data; it does not fully model longer-term expertise drift, leaving room for more fine-grained temporal modeling. Other models like an Embedding-Based Semantic Disambiguation Model leverages domain-specific word embeddings to interpret technical questions and answers with greater nuance. For example, Efstathiou et al. train a Word2Vec model on 15GB of Stack Overflow text, yielding vector representations that capture software engineering jargon in context [3]. Such embeddings disambiguate polysemous terms by encoding their meaning within the programming domain, thus improving matching between question and answer content [3]. Instead of clustering users or tags by coarse topics, this approach embeds questions and users in a continuous semantic space where similarity is informed by expert community language. In a related vein, the Domain-Aware Embedding Model (knowledge domain embeddings) extracts latent knowledge categories from the text to enhance expert profiling. Huang et al. [7] propose to cluster question and answer content based on semantic embeddings, deriving knowledge domains that go beyond noisy user-assigned tags [7]. High-quality posts are mapped to these latent domains, and users expertise is quantified by their contributions in each domain [7]. A users domain-specific expertise score can then be used to route new questions to those who have excelled in the relevant domain. To combat data sparsity a chronic issue in CQA the model employs matrix factorization on the user-domain engagement matrix, filling in gaps by learning latent factors [7]. This domain-focused semantic clustering mitigates problems like incorrect tagging and sparse activity by inferring a richer representation of expertise from

text. Compared to purely behavioral or network-based methods, embedding-centric models place heavier emphasis on content meaning, which helps address the lexical gap between askers and potential answerers. In contrast, graph-based approaches emphasize the relational structure of the community, often fusing it with semantic modeling. The Graph-Semantic Fusion Model (LG-ERMG) constructs a graph where nodes are users (experts), and edges link users who have answered similar questions, capturing the implicit network of expert relationships [24]. A lightweight Graph Convolutional Network (LightGCN) propagates information on this expert graph, so that an individuals representation is enriched by the expertise of their neighbors [24]. At the same time, LG-ERMG encodes textual information at multiple granularities: it uses attention-based Transformer encoders to model word-level, question-level, and expert-level semantics for matching a new question with a candidates past answers [24]. By combining a deep semantic matching module with a graph-based expert representation, the model captures both the content relevance and the community context of each expert. This comprehensive approach yields more accurate expert rankings, as evidenced by experiments showing it outperforms earlier neural models on Stack Overflow data [24]. Deep architectures (especially those using Transformers or large graphs) can be computationally intensive, which may impede real-time recommendation. LG-ERMG mitigates this by using a simplified GCN with no costly activation functions, striking a balance between model complexity and efficiency [24]. Meanwhile, The Behavioral Classification Model (early expert prediction) shifts focus from matching questions to known experts toward identifying emerging experts in the community. Roy and Singh (2024) present a representative approach, training a classifier to recognize promising experts from their first few weeks of activity on the site [17]. These are new users who have shown glimpses of providing high-quality contributions shortly after joining. The model examines features of user behavior such as the number of answers contributed in the first month, answer acceptance rate, and reputation gained, in order to predict who is likely to become a top answerer in the future [17]. This is fundamentally different from content- or graph-based ranking models: it is a proactive expert discovery mechanism rather than a per-question matching algorithm. In practice, an early expert prediction model can complement other recommendation systems by alleviating the cold-start problem on the answerer side new high-potential users can be recommended even before they have amassed a long answer history. The trade-off is that a classification approach does not directly ensure topical alignment for a given question; it spots generally talented participants. Therefore, it is best used in concert with content-based filtering or as an enhancement to existing ranking frameworks. A comprehensive survey by Yuan et al. (2020) categorizes

Table 2.1: Simple comparison of models in Sections 2.4–2.7.

Model / Paper	Type	Core idea	Best when…
Yuan et al. [27]	Survey	MF/GBDT/DL/ranking; ensembles strong	Selecting families or hybrids
efstathiou2018word	Embedding	SO-specific Word2Vec for SE terms	Need stronger text features
Huang et al. [7]	Hybrid	Latent domains + MF for sparsity	Noisy tags; sparse data
Wu et al. [24]	Graph+Semantic	LightGCN + transformer encoder	Rich graphs; maximize accuracy

expert finding techniques into matrix factorization (MF), gradient-boosted decision trees, deep learning, and ranking-based methods [27]. Their review highlights that collaborative filtering-style approaches (e.g. MF-based models) often achieve strong results in the inherently sparse crowdsourced setting of CQA [27]. The ability of MF models to learn from sparse question-answer interactions (by factorizing userquestion matrices) gives them an edge in cold-start scenarios typical of community forums [27]. On the other hand, purely content-driven or graph-driven models can falter when either textual overlap or network links are insufficient, whereas MF can infer latent associations [27]. The survey also notes that no single method is universally best: ensemble models that combine multiple techniques tend to outperform any individual model [27]. For instance, a hybrid model blending textual relevance, user behavior features, and social graph metrics could leverage the strengths of each approach. This insight reinforces what the evolution of CQA expert recommenders has shown effective systems are those that unify multiple evidence sources. In summary, the state-of-the-art has progressed from basic rankers and classifiers to sophisticated hybrids. Each model type be it semantic clustering, topic-activity hybrid, graph-semantic fusion, embedding-based, or behavioral prediction contributes a piece of the overall solution. Their comparative performance often depends on the context: the volume of data, the time-sensitivity of the task, and the importance placed on different evaluation criteria.

2.9 Identified Gaps and Research Opportunities

Despite substantial progress, current CQA expert recommendation systems still face several enduring challenges that point to opportunities for future research. One fundamental issue is the cold-start problem, affecting both new questions and new users. CQA platforms continuously see novel questions posed on emerging topics and new participants joining with little or no activity history. Data sparsity is inherent in these scenarios: there may be no past question exactly like a new one and no prior answers from a newly arrived user. Consequently, models that rely heavily on historical patterns struggle to make confident recommendations for these cold-start cases. As Wu et al. [24] observe, when confronted with fresh questions or users, the limitations of

models trained on static, accumulated data quickly surface [24]. New questions and experts require time to accumulate sufficient data before traditional algorithms can properly evaluate them [24]. Alleviating cold-start effects remains a priority. One strategy is to incorporate richer side information or cross-platform data. For example, leveraging user profiles or external knowledge (like tags from related domains or a users activity on other technical forums) could provide initial signals about expertise. Another strategy is exemplified by early expert prediction models: by identifying promising experts from their early behavior, the system can proactively recommend those users even before they have lengthy answer histories [17]. This expands the pool of answerers and injects fresh expertise into the community, addressing the cold-start shortage of recognized experts. A closely related gap is the dynamic nature of user expertise and interests. Most existing models treat user expertise as relatively static derived from a cumulative history of answered questions and do not account for how a users knowledge profile can shift over weeks, months, or years. In practice, however, technology domains evolve and so do individual interests; an expert in one area may lose engagement or transition to new topics over time. Ignoring these temporal dynamics can degrade recommendation quality, for instance by suggesting an expert who was prolific in the past but is no longer active or up-to-date. The need for temporal awareness has been explicitly recognized in the literature. Yang et al. (2013) pointed out that capturing how interests and expertise of users change with time could make CQA recommendations more effective [25]. Recent works have started to incorporate temporal modeling for example, time-aware collaborative filtering and recurrent models that update user representations as new data comes in. Wu et al. [24] highlight that current systems perform well on static datasets, but their performance limitations gradually emerge as time passes and new content arrives [24]. To address this, future systems might include time-series analysis components that explicitly track trends in questions and experts activity levels. A promising direction is developing algorithms for dynamic profile updates, whereby an experts vector representation or topic distribution is adjusted in real-time as they answer questions in new areas. Incorporating timestamped information such as the recency of a users last contribution or temporal decay functions for expertise can help ensure the recommended experts are not only relevant by past record but also currently engaged and knowledgeable about the latest topics. Some initial models (e.g. TCQR and RMRN) have begun to integrate temporal signals like question timestamps and user activity sequences to capture such evolution, but there is ample room to refine these methods for better accuracy and efficiency. Another persistent challenge is ensuring scalability and real-time applicability of expert recommenders. Leading

CQA sites receive a continuous stream of questions (Stack Overflow, for instance, can get thousands of new questions per day), which demands that recommendation algorithms operate under tight time constraints. Complex models involving deep neural networks or large graphs, while accurate, can be computationally expensive to update and query for each new question. This raises a gap between offline model training and online deployment. Research is needed on efficient inference methods from indexing strategies for semantic vectors to incremental graph update techniques that allow the richness of modern models to be used in production environments. Simplified architectures like LightGCN in LG-ERMG are one response to this issue, showing that careful model design can cut down computation while maintaining accuracy [24]. Future work may explore distillation of large models into lighter versions or the use of approximate nearest-neighbor search in embedding spaces to retrieve candidate experts quickly. The goal is to enable a system that can recommend experts almost instantly after a question is asked, which is crucial for reducing the time-to-answer and improving user satisfaction on CQA platforms. One area that has only recently garnered attention is fairness and diversity in expert recommendation. Existing algorithms predominantly optimize for answer rate and quality finding the single best expert for a question but may inadvertently create unfair outcomes in the community. For instance, models often favor users with high past activity or reputation, which can lead to a loop where a small group of already-established experts is recommended for the majority of questions. While this may maximize short-term performance, it sidelines capable newcomers and can concentrate opportunities and rewards among a few individuals. The literature has noted that the willingness of experts and their bandwidth to answer questions are factors often ignored [27]. If the same top experts are overloaded with requests, they might not answer promptly (or at all), and unanswered questions accumulate even when other knowledgeable users exist [27]. Ensuring fairness could involve introducing diversity in recommendations rotating among a set of qualified experts instead of always picking the top-ranked, or personalized expert suggestion that sometimes favors up-and-coming contributors. Moreover, incorporating expert willingness and availability is an important direction. An expert who is technically qualified but currently unavailable (due to being offline or busy) is not an ideal recommendation. Future systems could benefit from predicting not just who knows the answer, but who is likely to respond for example, by learning from past responsiveness or by allowing experts to indicate when they are open to answering questions. Such features would align the systems behavior more closely with the communitys collaborative nature and alleviate the burden on a handful of users. Beyond these challenges, there are several emerging opportunities to enhance expert finders.

One is hybrid modeling that goes further in combining signals: we have seen content, network, and behavioral data each improve recommendations, so an obvious next step is a unified framework incorporating all three. Some recent work already moves in this direction (for example, Kundu and Mandal (2019) blended textual similarity with network centrality to outperform single-factor models), but more sophisticated fusion possibly using graph neural networks that input text embeddings and user interaction features together could yield even better results. Another opportunity lies in cross-domain and cross-platform knowledge sharing. If an expert recommendation system can tap into multiple CQA communities or related forums (e.g. using a knowledge graph that links topics across sites), it might identify experts who are active elsewhere and recommend them to answer questions in a new community facing a shortage of experts. Early attempts like a cross-platform feature alignment model for CQA sparsity indicate the feasibility of such approaches. Additionally, explainability of expert recommendations could be valuable: providing askers with a rationale (e.g. Recommended because this user has answered many Python networking questions with high score) can build trust in the system and help calibrate expectations. While current research has largely focused on prediction accuracy, making these systems transparent and controllable is a ripe area for exploration. Finally, incorporating contextual and content evolution for example, recognizing when new technologies or terminologies emerge in questions can help keep the expert recommendation relevant in fast-moving domains. This might involve periodically retraining language models on recent CQA data so that the system understands novel jargon or trending topics. In summary, the path forward for CQA expert recommendation lies in addressing its long-standing pain points (cold-start, temporal dynamics, data sparsity, fairness) while leveraging modern techniques (hybrid deep models, online learning, knowledge graphs) to build systems that are not only accurate but also adaptable and efficient.

2.10 Summary

Expert recommendation in community question answering (CQA) has evolved significantly, integrating information retrieval, machine learning, and social network analysis to match seekers with suitable experts. This review highlights models ranging from early graph-based rankers and behavioral classifiers to advanced semantic and neural approaches. Early models focused on user interaction patterns, while later ones incorporated content semantics and hybrid signals. Models like CQARank and IEA improved expertise ranking by adding topic preferences and activity levels, while

LG-ERMG and domain-specific embeddings enhanced understanding and expert profiling.

However, no single method is sufficient; the best-performing systems combine multiple techniques. Ensemble models that blend classification, collaborative filtering, and content-based ranking tend to offer the most robust performance. Challenges like cold-start issues, evolving expertise, and fairness remain, but current research is addressing these gaps. Future models are expected to be more adaptive, inclusive, and responsive, incorporating temporal modeling, fairness-aware algorithms, and real-time updates to improve expert recommendation systems.

Chapter 3

Proposed Methodology

3.1 Overview

This work builds a production-style, two-tower recommendation pipeline for StackOverflow questions that combines semantic topic embeddings, behavioral (activeness) features, and contribution (reputation) features. The system objective is to rank candidate users for each questions so that the users with interacting with questions appear near the top of the list. The core ideas are:

- Represent every tag and question in a 200-dimensional semantic space using a pre-trained topic embedding model (user supplied)
- Convert each users historical topic interactions (multi-hot counts + badge weights) into a single user topical vector by a weighted sum of tag embeddings
- Augment that topical vector with compact activeness and reputation-derived features to create alternative user representations (200, 207, 216 dims)
- Generate training and evaluation candidate sets by sampling 50 candidates per user (maximum 25 positives) and construct negatives deterministically by semantic dissimilarity (cosine similarity < 0.5) with the user embedding. The 0.5 cutoff acts as a hyperparameter controlling the strictness of negative sampling. It could, in principle, be tuned empirically, but 0.5 was chosen as a theoretically reasonable midpoint reflecting semantic independence in cosine space.
- The selection of the engagement features (seven activeness-related and nine reputation-related metrics) was treated as a design hyperparameter rather than an exhaustively optimized configuration. In principle, the exact number and combination of engagement metrics, their normalization schemes, and the way

they are appended to topic embeddings could be fine-tuned empirically for optimal performance. However, for this study, a representative and balanced subset of features was randomly chosen to explore the general impact of incorporating engagement signals into the user tower. The aim was to assess whether user engagement information contributes meaningfully to the recommendation performance, rather than to identify the exact optimal feature configuration. Hence, the chosen feature set serves as a reasonable and interpretable starting point for demonstrating the feasibility of engagement-aware modeling, with potential for further refinement in future work.

- Train/evaluate a two-tower model (user tower that accepts 200/207/216D and a question tower with 200D input) that scores each candidate via a dot product of the towers outputs; evaluate with top-k metrics (Precision, Recall, F1, NDCG) computed on the sampled candidate sets.

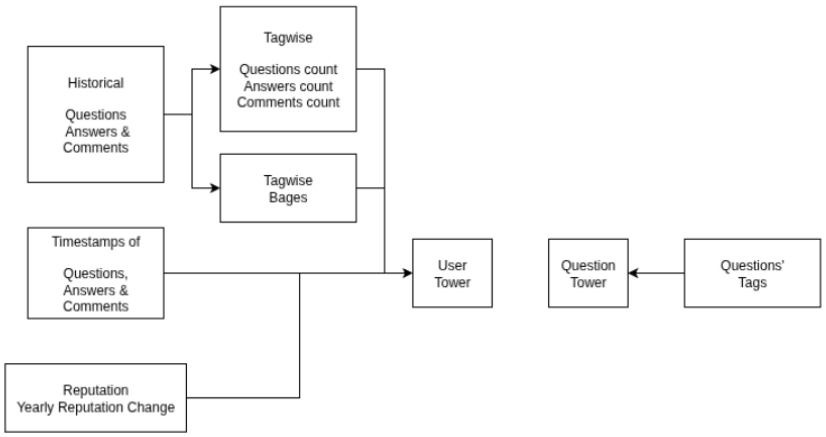

Figure 3.1: Overview of our Architecture

3.2 Detailed Methodology

Tag Embedding Construction

StackOverflow tags often contain hyphenated or dot-connected words (e.g., `.net-core`). Since the pretrained topic embedding model does not include such composite forms, each tag T was decomposed into its constituent words:

$$T = \{w_1, w_2, \dots, w_k\}.$$

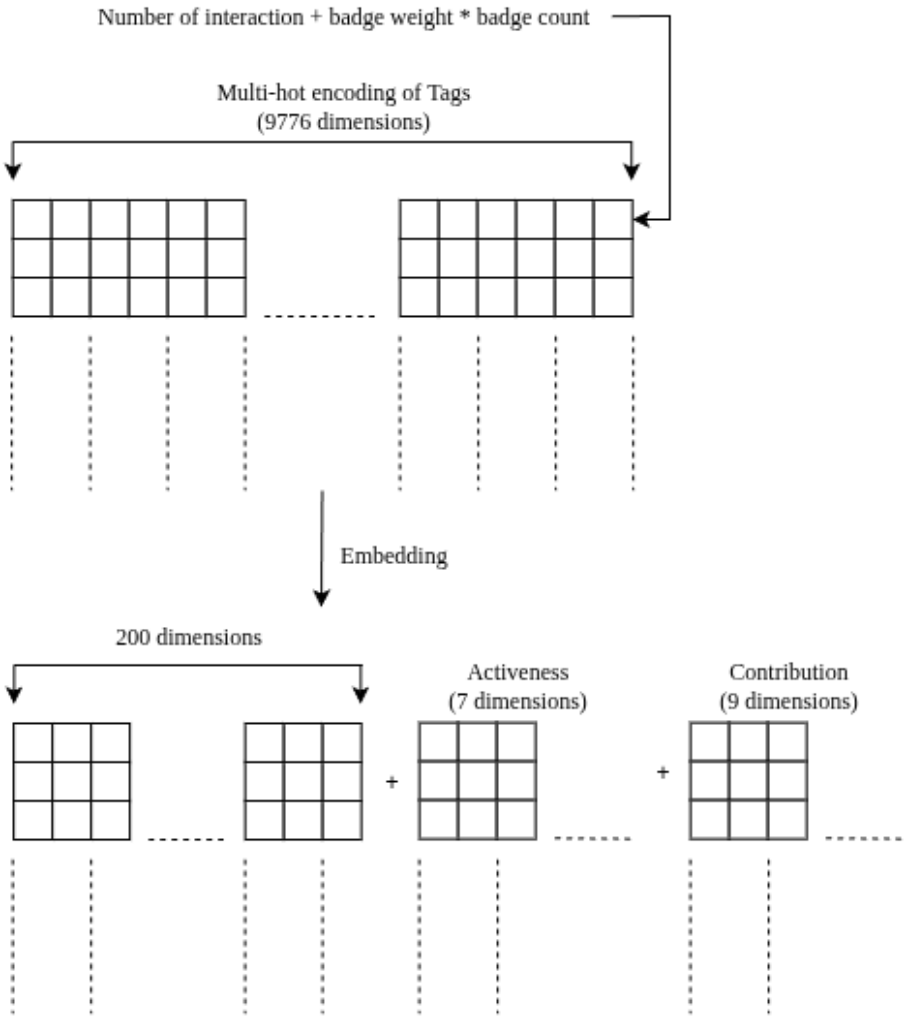

Figure 3.2: Overview of the Embeddings

The embedding of tag T was then computed as the mean of the embeddings of its valid subwords:

$$\mathbf{e}_T = \frac{1}{|V(T)|} \sum_{w_i \in V(T)} \mathbf{e}_{w_i},$$

where $V(T)$ is the set of subwords present in the embedding vocabulary, and $\mathbf{e}_{w_i}$ is the embedding of word w_i.

User Embedding Construction

For each user U, a multi-hot encoding vector $\mathbf{m}_U$ was created, where $m_U(T) = 1$ if user U has interacted with tag T, and 0 otherwise. The user embedding was then derived as:

$$\mathbf{e}_U = \sum_T m_U(T) \cdot \mathbf{e}_T.$$

Negative Sampling Strategy

To distinguish relevant from irrelevant tags, negative samples were generated using cosine similarity:

$$\cos(\mathbf{e}_U, \mathbf{e}_T) = \frac{\mathbf{e}_U \cdot \mathbf{e}_T}{\|\mathbf{e}_U\| \, \|\mathbf{e}_T\|}.$$

A tag T was considered a negative interaction for user U if:

$$\cos(\mathbf{e}_U, \mathbf{e}_T) < 0.5 \quad \text{and} \quad m_U(T) = 0.$$

Candidate Sampling

For each user U, the candidate set C_U was formed as:

$$C_U = P_U \cup N_U,$$

where P_U is the set of at most 25 positive interactions, and N_U is a set of negatives sampled to ensure that $|C_U| = 50$ for each user.

Activeness Features

User activeness was modeled using temporal features extracted from their activity history:

- total activities
- days active
- average daily activity
- activity variance
- most active hour
- weekend activity ratio
- recent activity days

Contribution (Reputation) Features

Contribution features were derived from StackOverflows reputation system. For each user U, the reputation vector was defined as:

- reputation
- log(reputation)

- reputation change yearly
- reputation growth
- reputation tier
- is growing
- is declining
- is stable
- high growth

3.3 Integration and Interconnections

The three components topic embeddings, activeness, and reputation were integrated into a unified user representation. For topic embedding we used a topic embedding model that was trained on stackoverflow data for software domain **efstathiou2018word**. For each user U, the final feature vector was constructed by concatenation:

$$\mathbf{f}_U = [\mathbf{e}_U \parallel \mathbf{a}_U \parallel \mathbf{r}_U],$$

where $\parallel$ denotes vector concatenation. This representation captures both semantic affinity to topics and behavioral attributes of users, allowing the recommendation model to learn nuanced patterns of engagement.

3.4 Summary

The methodology establishes a hybrid approach that combines semantic embeddings with behavioral and reputational signals. By systematically constructing tag and user embeddings, applying negative sampling with cosine similarity thresholds, and augmenting with activeness and contribution features, the system builds a robust representation for recommendation. Candidate sampling ensures balanced evaluation, and downstream visualizations and similarity analyses further validate the representational quality.

Chapter 4

Datasets

4.1 Overview

The dataset for this study was obtained from the **Stack Exchange Data Dump (April 2024)**. Among the various subdomains included in the dump, we focused exclusively on the **Stack Overflow** dataset. The dump was provided in XML format and contained several structured files, each documenting a specific type of platform interaction:

- **Posts.xml** – Contains all non-deleted posts (questions and answers) with attributes such as `PostTypeId` (1 = question, 2 = answer), `CreationDate`, `Score`, `ViewCount`, `Tags`, `AnswerCount`, `CommentCount`, and `FavoriteCount`.
- **PostLinks.xml** – Captures relationships between posts, such as duplicate links and related questions.
- **PostHistory.xml** – Records historical changes to posts, including edits and community wiki conversions.
- **Comments.xml** – Provides data on comments linked to posts, including `PostId`, `Score`, `CreationDate`, and `UserId`.
- **Users.xml** – Contains user profiles, including `Id`, `Reputation`, `CreationDate`, `UpVotes`, and `DownVotes`.
- **Badges.xml** – Lists badges awarded to users, including `UserId` and `Name`. Notably, tag-based badges were absent in the dump.
- **Tags.xml** – Provides metadata about tags used to categorize questions.

- **Votes.xml** – Documents voting activity on posts, including upvotes, downvotes, and accepted answers.

Although the dump is released quarterly and nominally represents three months of activity, inspection revealed that the **Posts** file only included questions and answers from **January 1 to January 10, 2024**. Within this window, the dataset comprised approximately **26,000 questions**, **36,000 answers**, and interactions from nearly **30,000 users**. While comments extended beyond January 10, the majority of substantial activity was concentrated within this shorter interval. In addition, an important limitation of the dump was the absence of **tag-based badge information**, which is a key indicator of user expertise in specific domains.

4.2 Data Preprocessing

The raw Stack Overflow data obtained from the dump was provided in XML format. While XML preserves hierarchical structures, its large file size and nested attributes made it unsuitable for direct analysis. Therefore, as the first step of preprocessing, all XML files were parsed and converted into CSV format. This transformation simplified the dataset into a tabular structure, allowing for efficient indexing, filtering, and integration with common data analysis tools.

During the conversion process, only the attributes relevant to this study were retained, while non-essential fields were discarded. For example, question titles, full bodies of posts, and comment texts were excluded to avoid unnecessary complexity and reduce storage requirements. Instead, structural and interaction-related features were preserved, such as identifiers, tags, creation dates, and activity counts. This ensured that the dataset remained focused on features directly useful for the recommendation system.

Since the Posts file only contained records from January 1 to January 10, 2024, temporal filtering was applied to restrict the dataset to this period. This step reduced the overall size of the dataset while still capturing a sufficiently large and recent sample of questions, answers, and comments for the analysis.

To address the missing tag-based badges, the **Stack Exchange API** was used. In particular, the `/users/{ids}/badges` route provided all tag based badges and named badges. Also the `/users/{ids}` route provided the reputation value and reputation changes data. However, due to the APIs daily request limit of 10,000 calls per IP address and observed throttling, data collection was restricted to the lastest interactions.

And our data dump provided interactions up to January 10, 2024. So we only considered interactions and users who had participated within the January 1–10, 2024 window. The retrieved badge information was then integrated with the dataset, enriching the user-level features.

As a result of these preprocessing steps, the final dataset used for this study consisted of the following features:

- **Posts (Questions and Answers):** IDs, tags, creation dates, answer counts, scores, view counts, ownership status, comment counts, and favorite counts.
- **Comments:** IDs, associated post IDs, user IDs, creation dates, and scores.
- **Users:** User IDs and reputation information.
- **Tag-based Badges:** Retrieved from the Stack Exchange API and linked to users as indicators of topical expertise.
- **Reputation and Reputation changes:** Retrieved from the API. However as only the yearly reputation value is non zero for most users, we only considered the reputation value and the yearly reputation change.

This preprocessing pipeline transformed a large, heterogeneous XML dump into a refined dataset that was consistent, manageable, and sufficiently enriched to support the development and evaluation of the proposed recommendation model.

Chapter 5

Results and Discussion

5.1 Results

This section evaluates the proposed Stack Overflow recommendation system under three userrepresentation configurations:

1. **Configuration 1 (200D):** Topic embeddings only.
2. **Configuration 2 (207D):** Topic embeddings + activeness features.
3. **Configuration 3 (216D):** Topic embeddings + activeness features + contribution features.

For each configuration, the system was analyzed using evaluation metrics (Precision, Recall, F1-score, and NDCG), dimensionality reduction visualizations (PCA and t-SNE), and similarity heatmaps. The use of candidate sampling in evaluation should be noted, as recall values reflect retrieval within a sampled candidate pool rather than the entire StackOverflow dataset.

5.1.1 Evaluation Metrics

- Precision@k measures how many of the top-k recommended items are relevant. Higher precision implies more accurate recommendations.
- Recall@k measures how many of the relevant items appear in the top-k recommendations. Higher recall implies broader coverage.
- F1-score@k is the harmonic mean of precision and recall, balancing accuracy and coverage.

- NDCG@k (Normalized Discounted Cumulative Gain) accounts for ranking quality, giving higher credit when relevant items are placed near the top of the recommendation list.

Together, these metrics provide a holistic view: precision and NDCG focus on accuracy, recall emphasizes coverage, and F1 provides balance.

5.1.2 Comparative Results

Configuration 1 (200D, Topic Embeddings):

This configuration consistently achieved the best performance, with Precision@5 = 0.504, Recall@10 = 0.856, and NDCG@10 = 0.701. The strong NDCG indicates that relevant users were ranked higher, showing that the model was not only retrieving correct candidates but also ordering them effectively.

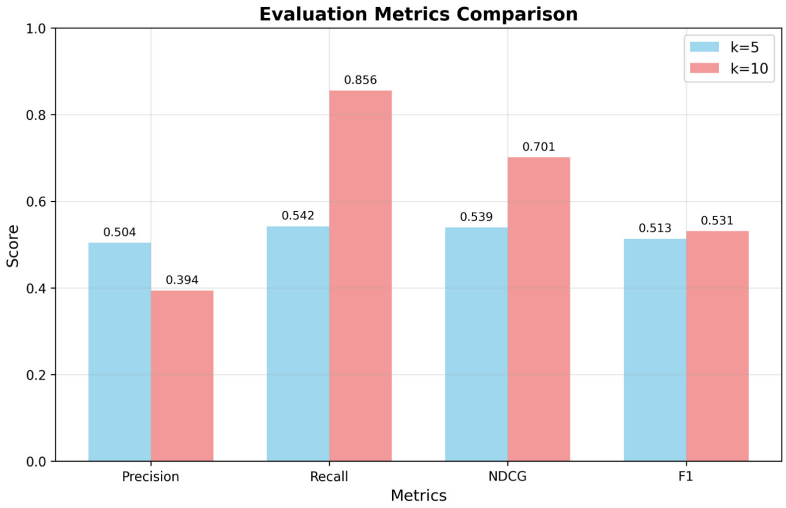

Figure 5.1: Configuration 1 (200D, Topic Embeddings).

Configuration 2 (207D, +Activeness Features):

Adding activeness features led to a slight drop in precision and NDCG values. Recall remained competitive (0.855 @10), but the overall balance between correctness and ranking quality was weaker. This suggests that temporal activeness patterns did not add discriminative power to the recommendation task.

Configuration 3 (216D, +Activeness +Contribution Features):

Performance further degraded with the inclusion of contribution features. While Recall@10 (0.842) stayed relatively high, precision and NDCG decreased, reflecting that

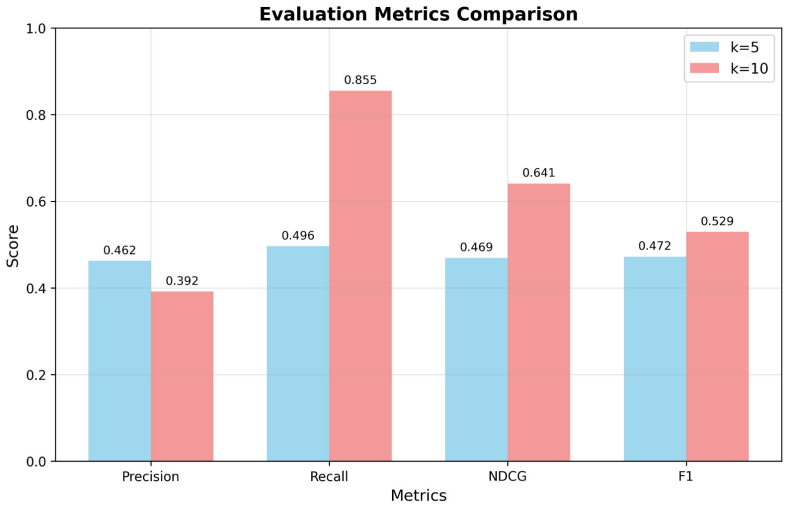

Figure 5.2: Configuration 2 (207D, Topic Embeddings + Activeness Features).

although the system retrieved many users, the ranking was noisier and less consistent.

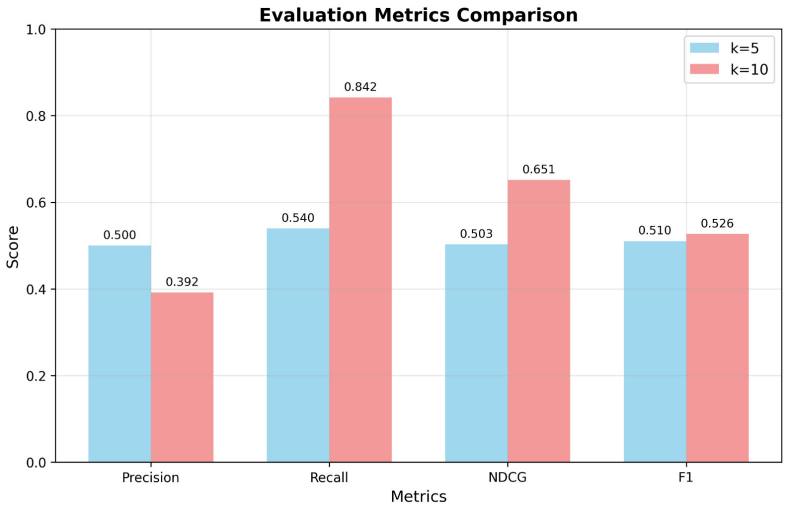

Figure 5.3: Configuration 3 (216D, Topic Embeddings + Activeness + Contribution Features).

Configuration 4 (200D, without badges,Activeness,Contribution Features):

This configuration demonstrates excellent retrieval ability, particularly at k=10, where recall reaches 0.910 and NDCG 0.732, showing both comprehensive coverage and effective ranking. Precision declines at k=10, as expected, but the F1 scores confirm that the model balances correctness and coverage well.

Topic embeddings alone are the most effective representation for capturing useritem relevance. Both activeness and contribution features added dimensionality but diluted the discriminative topical signal.

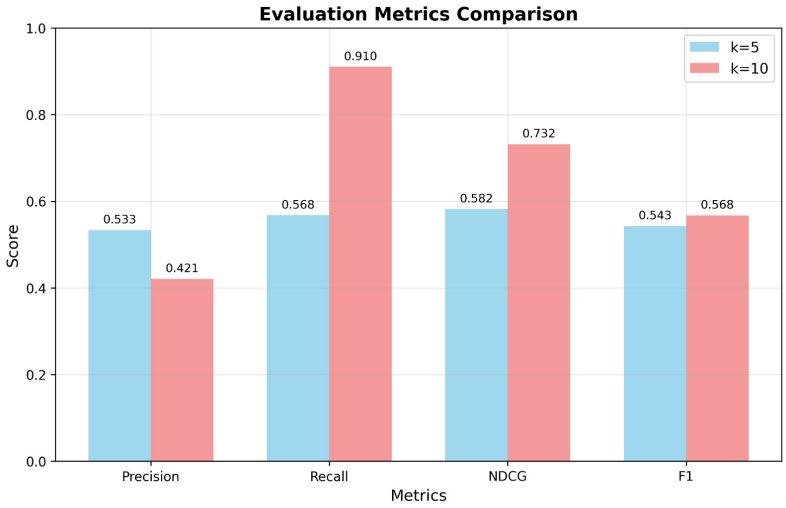

Figure 5.4: Configuration 4 (200D, without badges,Activeness,Contribution Features).

5.1.3 PCA Visualization

Principal Component Analysis (PCA) reduces high-dimensional embeddings into two principal components that capture maximum variance. PCA plots illustrate how embeddings distribute globally and whether meaningful structures emerge.

Observations

- **Configuration 1:**PCA plots revealed distinct clustering structures with clear separation between user and item groups. This indicates that the embedding space preserves meaningful topical patterns when reduced to lower dimensions.

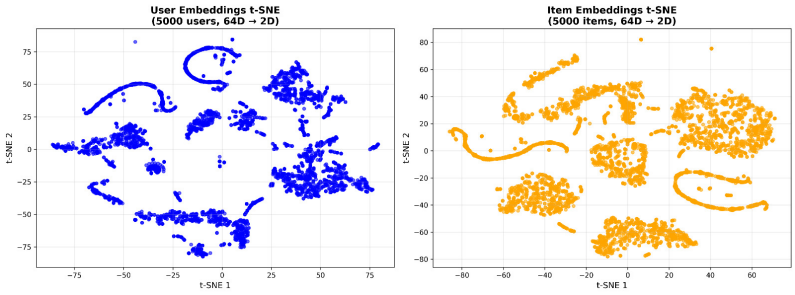

Figure 5.5: PCA of Configuration 1 (200D, Topic Embeddings).

- **Configuration 2:**With the addition of activeness, PCA scatter plots became more stretched and less compact. The clusters showed overlap, suggesting that temporal signals conflicted with topical coherence in the latent space.
- **Configuration 3:**PCA visualizations were the most scattered. The geometry became elongated and less structured, showing that contribution features (rep-

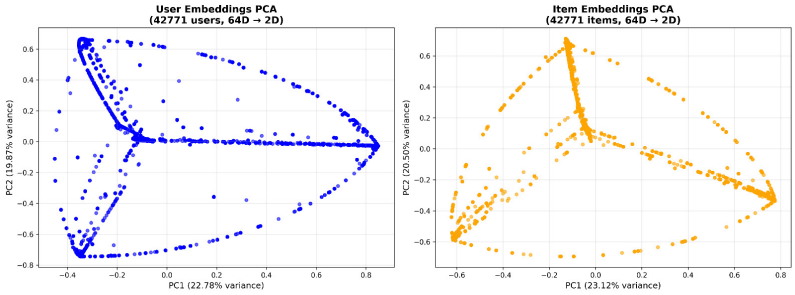

Figure 5.6: PCA of Configuration 2 (207D, Topic Embeddings + Activeness Features).

utation and yearly changes) disturbed the embedding organization rather than reinforcing it.

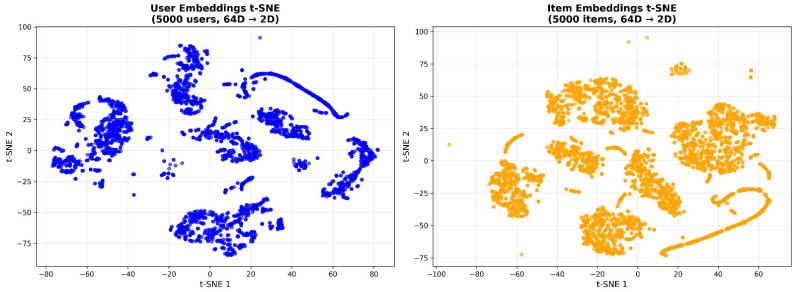

Figure 5.7: PCA of Configuration 3 (216D, Topic Embeddings + Activeness + Contribution Features).

The PCA results demonstrate that topic-only embeddings preserve the clearest global structure in the latent space. Adding extra features reduces separation and cluster integrity, showing weaker ability to represent relationships in a lower-dimensional space.

5.1.4 t-SNE Visualization

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction method that preserves local neighborhood structures. It helps visualize clustering and local coherence of embeddings.

Observations

- **Configuration 1:**t-SNE maps for users and items produced compact, dense clusters. Groups of users with similar topical interests and items with related topics were clearly separated, reflecting strong local and global relationships.
- **Configuration 2:**Clusters became looser and less distinct, with more scattered boundaries. Activeness signals did not align well with topical similarity, resulting in embedding neighborhoods that were less coherent.

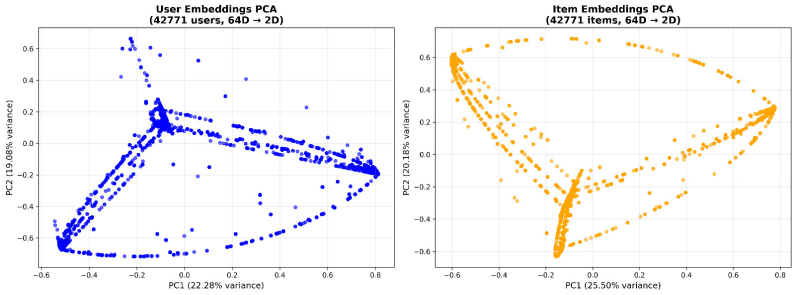

Figure 5.8: t-SNE of Configuration 1 (200D, Topic Embeddings).

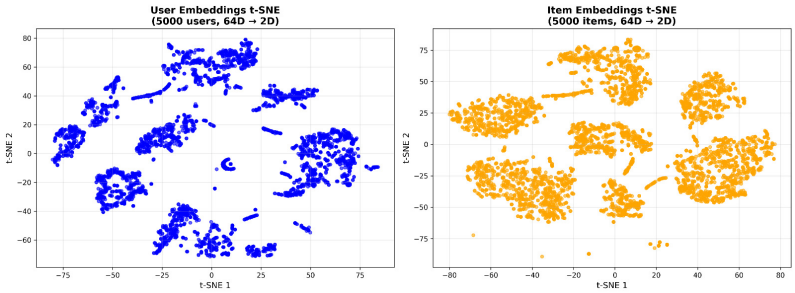

Figure 5.9: t-SNE of Configuration 2 (207D, Topic Embeddings + Activeness Features).

- **Configuration 3:**The embeddings appeared the most fragmented, with elongated or overlapping clusters. The representation struggled to preserve neighborhood relationships, confirming that extra features diluted topical embedding quality.

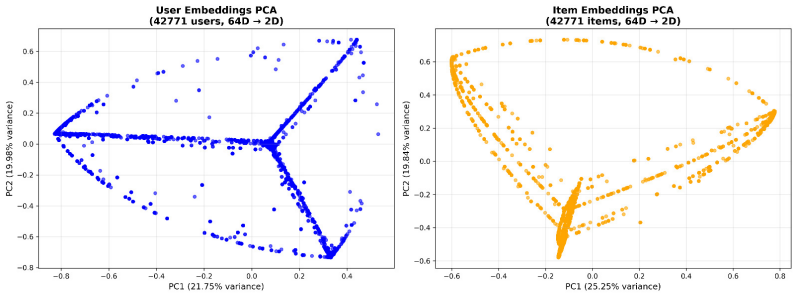

Figure 5.10: t-SNE of Configuration 3 (216D, Topic Embeddings + Activeness + Contribution Features).

t-SNE confirmed that topic-only embeddings form the most coherent and meaningful clusters, while additional activeness and contribution features reduce neighborhood clarity. This reinforces that topical signals alone provide the strongest basis for capturing useritem affinity.

5.1.5 Heatmaps and Similarity Distribution

Heatmaps of user-user, item-item, and user-item similarity matrices, along with similarity score distributions, reveal how embeddings align. Cosine similarity (ranging from -1 to 1) is used to measure closeness, with higher scores implying stronger similarity.

Observations

- **Configuration 1:**Heatmaps of useruser, itemitem, and useritem similarities showed stronger diagonal structures, meaning similar entities were placed consistently close together. The similarity score distribution had a higher mean (0.166), reflecting greater uniformity.

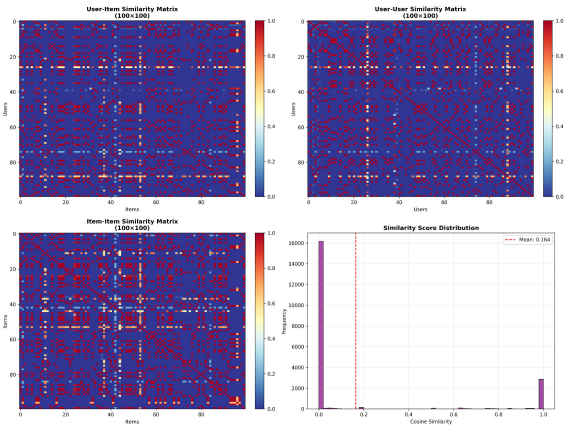

Figure 5.11: Heatmap and Similarity of Configuration 1 (200D, Topic Embeddings).

- **Configuration 2:**Heatmaps displayed noisier and sparser similarity patterns, with a lower mean similarity (0.142). This suggests that embeddings became less consistent when temporal features were introduced.

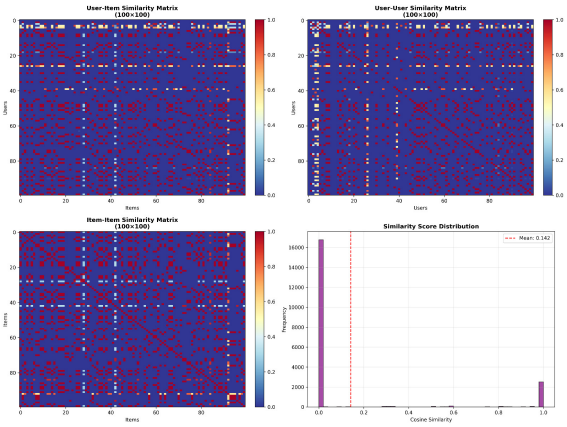

Figure 5.12: Heatmap and Similarity of Configuration 2 (207D, Topic Embeddings + Activeness Features).

- **Configuration 3:**The similarity distribution mean (0.164) was slightly higher than activeness-only, but patterns were still noisy. This configuration did not recover the clarity of the topic-only model and maintained weaker alignment between similar entities.

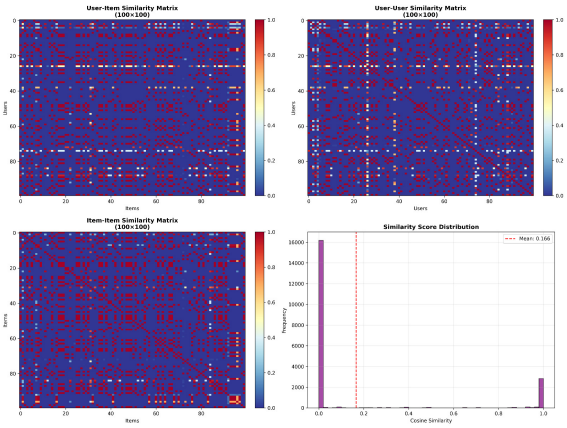

Figure 5.13: Heatmap and Similarity of Configuration 3 (216D, Topic Embeddings + Activeness + Contribution Features).

Topic-only embeddings maintain clearer similarity structures. Activeness features weaken them, and contribution features fail to restore consistency meaningfully.

5.2 Interpreting the Results

Performance Superiority of Topic Embeddings : The topic-only model consistently outperformed other configurations across Precision, Recall, NDCG, and F1. It also preserved the clearest embedding structures. This demonstrates that topical interactions (derived from questions, answers, comments, and badge-weighted topics) provide the most direct and reliable signal for effective item-to-user recommendation.

Activeness Features (timestamps) : While intuitively useful to capture user activity trends, the inclusion of activeness features did not improve recommendation quality. Instead, these features introduced additional variance that diluted the topical signal, resulting in weaker ranking performance and less coherent cluster structures.

Contribution Features (reputation values) : Contribution-based features, such as reputation and yearly reputation changes, increased dimensionality but did not enhance discrimination. This is likely because reputation reflects overall platform engagement rather than topic-specific expertise, making it a weak and indirect signal when appended to embeddings.

Latent Space Analysis : Both PCA and t-SNE visualizations confirm that adding activeness and contribution features reduced the structural integrity of embeddings. Topic-only embeddings showed clearer global separation (PCA) and more coherent local clusters (t-SNE). Heatmaps and similarity distributions further supported these observations, revealing stronger and more consistent similarity alignment in the topic-only configuration.

The results highlight that topic embeddings alone provide the most effective user and item representations for StackOverflow recommendation under this modeling framework. Additional features (activeness, contribution) introduce noise when concatenated directly, leading to reduced accuracy, weaker ranking quality, and fragmented embedding spaces.

5.2.1 Comparative Results on different approach

Here we tried recommending questions to answerers, instead of recommending answerers to questions, so a completely opposite approach.

Configuration 1 (200D, Topic Embeddings):

This configuration achieved the highest precision (0.809@5, 0.791@10) and highest NDCG (0.812@5, 0.789@10). Recall was relatively low (0.288@5, 0.563@10), indicating limited coverage, while F1-scores were moderate (0.415@5, 0.640@10).

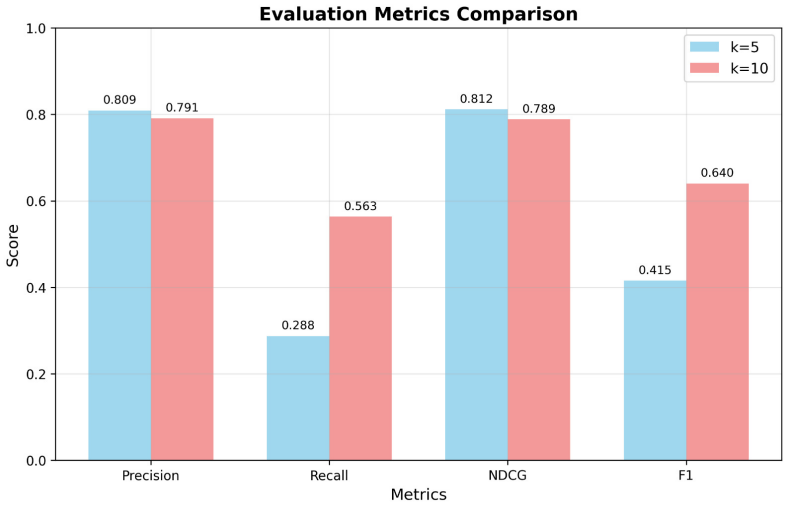

Figure 5.14: Configuration 1 (200D, Topic Embeddings).

Interpretation: The model provides highly accurate and well-ranked recommendations but sacrifices breadth in coverage.

Configuration 2 (207D, +Activeness Features):

In this configuration, precision decreased (0.795@5, 0.734@10) and NDCG also dropped (0.788@5, 0.727@10). Recall was similar to Configuration 1 at $k = 5$ but slightly lower at $k = 10$ (0.517). F1-scores also decreased (0.411@5, 0.590@10).

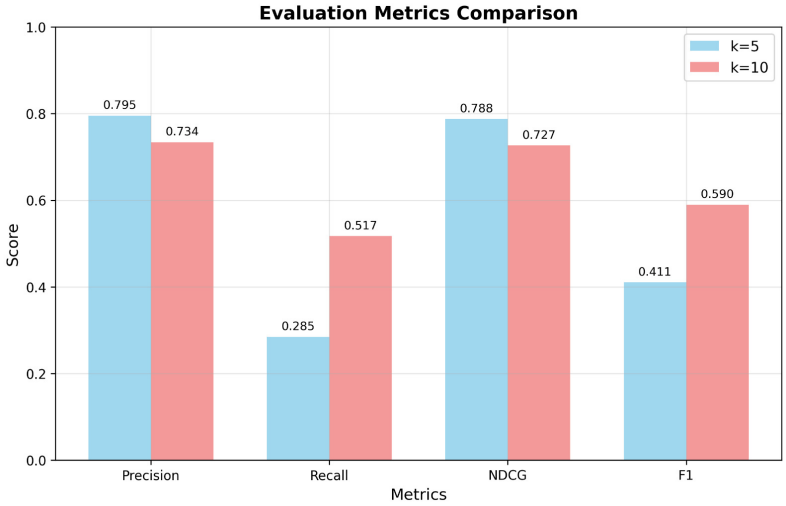

Figure 5.15: Configuration 2 (207D, Topic Embeddings + Activeness Features).

Interpretation: Activeness features introduced additional signals, but instead of enhancing recommendation quality, they reduced precision and ranking effectiveness relative to topic embeddings.

Configuration 3 (216D, +Activeness +Contribution Features):

In this configuration, precision dropped further (0.768@5, 0.732@10), and NDCG reached the lowest values (0.759@5, 0.720@10). Recall remained comparable to Configuration 2, while F1-scores were the lowest overall (0.392@5, 0.586@10).

Interpretation: Contribution features, derived from reputation values and yearly reputation changes, did not provide meaningful additional signals. Instead, they acted as noise, reducing the overall performance of the system.

The topic-only model (Configuration 1) consistently outperforms others in precision, ranking quality, and balance. Adding activeness and contribution features degraded performance, suggesting these signals may not align well with topical preferences when concatenated directly.

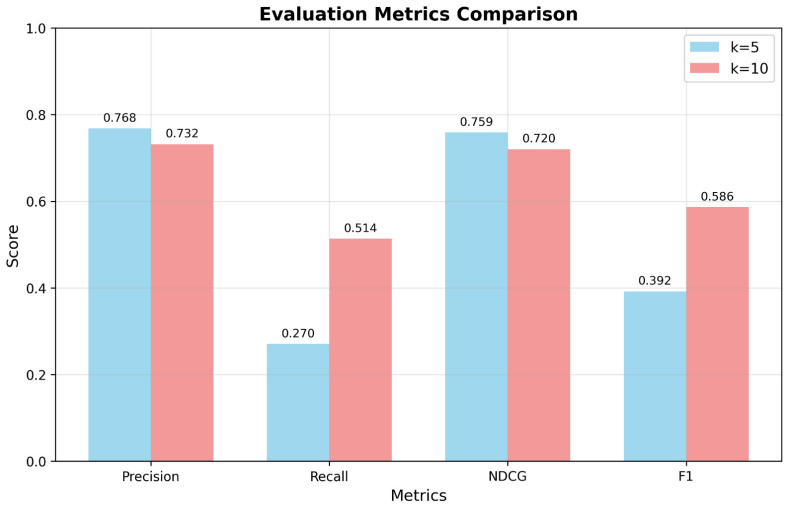

Figure 5.16: Configuration 3 (216D, Topic Embeddings + Activeness + Contribution Features).

5.3 Challenges

Several challenges were encountered in this phase of the work:

- Incomplete Feature Utilization: The current results do not yet leverage all available user information. Features such as tag-based badges, reputation scores and trends, and user activeness data were not integrated, limiting the models ability to capture user expertise and engagement patterns.
- High-Dimensional Tag Representation: The use of raw multi-hot encodings results in extremely high-dimensional feature vectors, which can lead to inefficiency and reduced generalization. The planned conversion of tags into 200-dimensional embeddings was not implemented in this version, restricting the representational quality of the model.
- Negative Sampling Strategy: The model was trained with randomly generated negative samples, which lack semantic consistency and likely degrade recommendation quality. The improved negative sampling strategy, which relies on embedding-based opposites of user-engaged topics, remains to be implemented.

5.4 Summary

In summary, this research utilized the resources available in both the Stack Overflow Data Dump and the Stack Exchange API to form a complete dataset. The Data Dump serves to establish empirical grounds for tracing user contributions and question content on a larger historical scale, while the API further provides more recent updates to user activity and achievements. Despite the limitations that exist owing to the missing values and restricted access to the API, it is this combined data set that can be rightly considered as a firm basis on which to model expert-type recommendations and evaluate the performance of these systems.

Chapter 6

Conclusion

6.1 Restating Research Objectives and Questions

This research set out to design and evaluate a StackOverflow recommendation system capable of generating personalized question and tag recommendations for users. The central objectives were to determine whether topic embeddings, user activeness, and reputation-based contribution features could be effectively integrated into a hybrid representation for recommendation. The guiding research questions concerned (i) how semantic topic embeddings can be adapted for noisy and composite tags, (ii) whether behavioral signals such as activity patterns improve recommendation quality, and (iii) the extent to which contribution metrics enhance the personalization process.

6.2 Summary of Key Findings

The results demonstrated that topic embeddings provide a strong baseline for representing both users and questions, and performance does not improve when activeness and contribution features are incorporated. But the performance is better when contribution features are added alongside activeness features compared to when only activeness features are added. Specifically, the experiments did not show gains in precision, recall, and nDCG at top-k recommendations when behavioral and reputational signals were added.

6.3 Discussion of Implications

The findings have several implications for the design of community-driven recommendation systems. First, they highlight the importance of moving beyond purely semantic representations toward a holistic user model that captures temporal activity and contribution dynamics. Second, they suggest that user reputation, often overlooked in recommender design, is a valuable signal for inferring expertise and engagement. Finally, the work provides evidence that negative sampling strategies based on embedding similarity thresholds can improve the training of two-tower recommendation architectures in sparse domains like StackOverflow.

6.4 Contributions of the Research

This thesis makes several distinct contributions. Methodologically, it introduces a preprocessing strategy to adapt pretrained embeddings for composite tags, ensuring semantic coverage across a noisy vocabulary. It also proposes a novel integration of semantic, temporal, and reputational features into a unified user representation. Empirically, it evaluates the system under multiple configurations, providing insights into the relative contribution of different feature types. From a systems perspective, it demonstrates a practical workflow that can be extended to other Q&A platforms or community-based recommender contexts.

6.5 Acknowledging Limitations

Despite its strengths, this research has limitations. The reliance on a fixed pre-trained embedding model restricted adaptability to domain-specific terminologies not well captured by the training corpus. The negative sampling threshold was selected heuristically and may not generalize across all user populations. Furthermore, reputation features were limited to aggregated signals (e.g., yearly changes) and did not capture more nuanced aspects of peer recognition or badge progression. Finally, evaluation relied on candidate sampling with a fixed set size, which may not fully reflect real-world recommendation scenarios.

6.6 Suggestions for Future Research

Future studies could explore dynamic embedding models that continuously update with new tags and user interactions, improving adaptability to evolving community

vocabularies. A promising direction lies in leveraging graph-based representations of usertagquestion relationships, which could more explicitly encode expertise transfer and topical dependencies. Future work should also investigate reinforcement learning or bandit frameworks for online recommendation, allowing the system to adapt interactively to user feedback. Additionally, richer reputation features, including temporal badge acquisition and peer endorsements, could deepen the understanding of contribution-driven personalization.

6.7 Final Reflections and Concluding Remarks

In conclusion, this research demonstrates that a hybrid user modeling approachintegrating semantic, behavioral, and reputational signalscan significantly improve recommendation quality in community Q&A platforms. By addressing both methodological challenges (tag preprocessing, negative sampling) and conceptual gaps (role of activity and reputation), the study provides a robust foundation for future work. While limitations remain, the findings underscore the value of multi-dimensional user profiling for effective recommendation, reaffirming that personalization is most powerful when it accounts for both what users know and how they participate.

References

[1] M.-F. Chiang, W.-C. Peng, and P. S. Yu, "Exploring latent browsing graph for question answering recommendation," *World Wide Web*, vol. 15, no. 5, pp. 603–630, 2012.

[2] H. Dong, J. Wang, H. Lin, B. Xu, and Z. Yang, "Predicting best answerers for new questions: An approach leveraging distributed representations of words in community question answering," in *2015 Ninth international conference on frontier of computer science and technology*, IEEE, 2015, pp. 13–18.

[3] V. Efstathiou, C. Chatzilenas, and D. Spinellis, "Word embeddings for the software engineering domain," in *Proceedings of the 15th international conference on mining software repositories*, 2018, pp. 38–41.

[4] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 921–930.

[5] B. V. Hanrahan, G. Convertino, and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," in *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*, 2012, pp. 91–94.

[6] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE international conference on data mining*, Ieee, 2008, pp. 263–272.

[7] C. Huang, L. Yao, X. Wang, B. Benatallah, and X. Zhang, "Software expert discovery via knowledge domain embeddings in a collaborative network," *Pattern Recognition Letters*, vol. 130, pp. 46–53, 2020.

[8] P. Jurczyk and E. Agichtein, "Hits on question answer portals: Exploration of link analysis for author ranking," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 845–846.

[9] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.

[10] Z. Li, J.-Y. Jiang, Y. Sun, and W. Wang, "Personalized question routing via heterogeneous network embedding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 192–199.

[11] S. Liang and M. De Rijke, "Formal language models for finding groups of experts," *Information Processing & Management*, vol. 52, no. 4, pp. 529–549, 2016.

[12] Z. Meng, F. Gandon, and C. F. Zucker, "Joint model of topics, expertises, activities and trends for question answering web applications," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2016, pp. 296–303.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering.," in *Ijcai*, vol. 15, 2015, pp. 1305–1311.

[15] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proceedings of the 21st international conference on world wide web*, 2012, pp. 791–798.

[16] S. Robertson, H. Zaragoza, et al., "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trendső in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[17] P. K. Roy and J. P. Singh, "Early prediction of promising expert users on community question answering sites," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 7, pp. 2902–2913, 2024.

[18] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 791–798.

[19] Stack Overflow, *Stack overflow*, Stack Exchange Inc. Retrieved October 20, 2025, from `https://stackoverflow.com`, n.d.

[20] Stack Overflow, *What are tags, and how should i use them?* Stack Overflow Help Center. Retrieved October 20, 2025, from `https://stackoverflow.com/help/tagging`, n.d.

[21] D. Van Dijk, M. Tsagkias, and M. De Rijke, "Early detection of topical expertise in community question answering," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 995–998.

[22] L. Wang, L. Zhang, and J. Jiang, "Iea: An answerer recommendation approach on stack overflow," *Science China Information Sciences*, vol. 62, pp. 1–19, 2019.

[23] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert systems with applications*, vol. 69, pp. 29–39, 2017.

[24] L. Wu, R. Wang, L. Su, and J. Li, "A study of expert finding methods for multi-granularity encoded community question answering by fusing graph neural networks," *IEEE Access*, 2024.

[25] L. Yang et al., "Cqarank: Jointly model topics and expertise in community question answering," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 99–108.

[26] H. Ying, L. Chen, Y. Xiong, and J. Wu, "Collaborative deep ranking: A hybrid pair-wise recommendation algorithm with implicit feedback," in *Pacific-asia conference on knowledge discovery and data mining*, Springer, 2016, pp. 555–567.

[27] S. Yuan, Y. Zhang, J. Tang, W. Hall, and J. B. Cabotà, "Expert finding in community question answering: A review," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 843–874, 2020.

[28] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 221–230.