Answer Selection In Community Question Answering Portals

Juee Gosavi
MIT College of Engineering, Pune, India
jueeg1693@gmail.com

B. N. Jagdale
MIT College of India, Pune, India
bjagdale@gmail.com

***Abstract*—New level of information sharing is enabled by different online communities like wikis , blogs, forums etc. which provides a platform for interaction with individuals which offers services like searching and posting queries or answers and share expertise with other information seekers. For recently posted query or searched query system furnishes the pool of answers with similar questions links, which could be a prolonged task for finding the significant answer. To overcome this, the system proposes an approach to effectively rank answers which are most relevant and best from historical archives based on similar queries found. It comprises of two components, one which contains training samples with positive, negative and neutral classes and other component retrieves similar questions to posted questions which are with their answer pools. Two data mining approaches were compared to retrieve similar questions. Our objective is to rank answer candidates based on pair wise comparison where question-answer pairs are ranked using pair wise learning to a rank approach based on a trained model which provides the user with most relevant answers for a given posted question.**

Keywords—Answer Selection, Community Question Answering, Ranking, Question Classification, Data Mining Algorithms

I. INTRODUCTION

Public Forums and Questions Answering sites have seen a spectacular increase in acceptance in the recent years. With the advent and popularity of sites like Yahoo! Answers, Cross Validated, Stack Overflow, Quora, Health Tap more and more people now use these web forums to get answers to their questions. These forums offer individuals the flexibility to post their queries online and have multiple experts across the world to answer them, whereas having the ability to produce opinions or expertise to help other users, a quality of answers encourages more participation and recognition. The leading form of knowledge retrieval is represented by question answering system which is recognized by user information requirements that are relatively conveyed in terms of natural language sentences or queries and is a type of natural sorts of human-computer interaction. Apparently with conventional data recovery, in question answering where whole contents are viewed as relevant to the asked data, as a reply to a query precise parts of data are returned. Finding a brief and comprehensible with correct answer, which refers to passage, word, picture, sentence, an audio fragment, or whole document, is required by users of a question answering system [9].

Online social networks growing rapidly nowadays which offers a wide range of options to express opinions in natural language. Question Answering Systems are one of the web forums which allow users or experts to ask or to answer to a question in natural language. The main functionality of the question answering system is, for a given question from web and collection of documents, find the exact answer which satisfies the user [8]. Multiple ways are present to find answers which are relevant to the question asked but for a particular query, long lists of probably relevant documents are returned by current community question answering sites without identifying the vital of the result with a brief answer. Therefore the most essential tasks for knowledge consumers or users is to identify precise answer data, which is getting a direct specific relevant answer for a query. Proper organization of required knowledge is needed by returning exact and specific answers [10].

In Community question and answer systems when we try to find answers to the questions we use archives where we can find them using theoretical base. But it can be time-consuming part to find out questions and where they can be associated with different answers and to find out relevant answers they need to go through a lot of answers to find what is needed [1]. It is necessary to find out a precise answer for a given query which most relevant and recent. Also, information seekers need to wait for a long time for receiving an answer from other users, so finding similar questions and answers from historical documents will help to reduce time. A social network that presents an option to conventional web findings is recognized as Community Question-Answering (CQA) forums. Users of forums enter their required information as a proper question in natural language and get straight replies written by people or experts, instead of retrieving results of web search networks. Natural language contents come in various qualities like questions-answers scopes from supreme quality content to low grade content to unrelated content or even offensive content. Due to which complications are increased in voting given to best answers and selection of supreme quality answer becomes most important [6].

The Question Answering system contains three methods as Question Classification (QC) which is a machine learning classifier, used for identification of the type of answer related to the input query based on training. The type of answer helps to identify a corresponding context-ranking model for ranking retrieved documents, Document and Passage gathered in their archives, which encourages the perpetuation and searching answered queries [10].

The remaining tpoics of this paper are arranged as shown below: section 2 represents the review and state of the literature. In section 3, we have design proposed system. Section 4 represents the algorithm used in our system. In

section 5, some implementation screenshots are shown and next section 6 includes the results and analysis of this system. Finally, last section 7, contains the conclusion of the paper.

II. Review Of Literature

In community question and answer systems, we have a tendency to realize answers we use archives where we can realize them using theoretical base. However, it can be time-consuming part to seek out queries and where they can be associated with different answers and to find out relevant answers they need to go through a lot of answers to find what is required. To overcome this problem following papers were referred:

The framework used in the paper [2] is based on the expertise level of the answerer in CQA session using SVM-based and Ranking SVM-based methods, questions are ranked and routed to the appropriate answerer. In results, Ranking SVM-based methods perform better than SVM-based methods on real-world datasets. The author in [3] proposed an approach based on TF-IDF which identifies the person who answers best for a recently asked question which contains vector space model where user's interest and user's expertise are considered while selecting best answerer. Based on historical question archives using Latent Dirichlet Allocation (LDA) answerers interests are modeled.

Paper [4] shows the empirical study and analysis of the answers to predict acceptability of the answers by the asker or community user by using Bayes classifier model, the approach identifies topic modeling to extract features for pattern identification for selecting best answers. This approach analyzes Stack Overflow Q&A for selecting best answer whereas method in [5] represents an automatic content migration from web forums to latest Question answering forums based on binary classifier built upon text features for identifying best answers with good performance. Results provide a positive approach against automatic migration of crowd sourced information from legacy forums to latest question answering sites.

In another study [6], the system uses a perceptron and a ranking Support Vector Machine based method of the stochastic gradient descent frame work, for regularization and learning from noisy data for the improvement in the answer ranking for social QA with the use of feature engineering along with learning procedures. The outcome indicates the helpfulness of query expansion strategies as well as the effect of regularization at the time of learning from noisy data.

W. Wei in [10] proposed a three-level scheme which is based on generating a query-oriented summary format answer in form of novelty and redundancy. It calculates the global ranking score and combines it with a relevance score. It is based on calculated global ranking scores, which uses two various methods to build top K no. of answer set, and then solves an optimization problem to generate as a summary of top answers to a question asked by user. P. Roy in study [9] proposed a method to rank responses depending on their quality and new tab is introduced as promising answers. First feature extraction is applied on dataset and then gradient boosting classifier, naïve bayes and random forest and is applied to train this system. The answers are classified into high quality, medium quality and low quality answers.

III. Proposed System

The Main idea of this system is to effectively rank answers which are most relevant and best from historical archives based on similar queries found. The system architecture comprises of modules which are an offline module and an online module which utilizes K nearest neighbor algorithm compared with Naïve Bayes algorithm for similar question retrieval and SVM-based rank model for finding most relevant answers for the newly searched question in which real-time HealthTap.com dataset is utilized as a part of training and testing data. The dataset comprises of queries asked by various patients which are replied by various specialists and user ratings given by different users of HealthTap.com community. The precision and recall values are compared for retrieving similar questions along with most relevant answers.

A. Online Module :

In online component user searches or posts any question, first it preprocesses the question and then it searches similar question and answers from repositories.

Preprocessing Module: Here, the asked question is processed by removing stop words and keyword extraction. Natural language processing is applied to process question and tag question features using part of speech tagging.

Extracting relevant questions and their answers: Based on the syntactic relevance of similar questions with their answers are retrieved from the database. For this process, KNN strategy is used to find top similar questions Also Naïve Bayes is performed to retrieve relevant questions to posted query or searched query and to compare results with KNN algorithm.

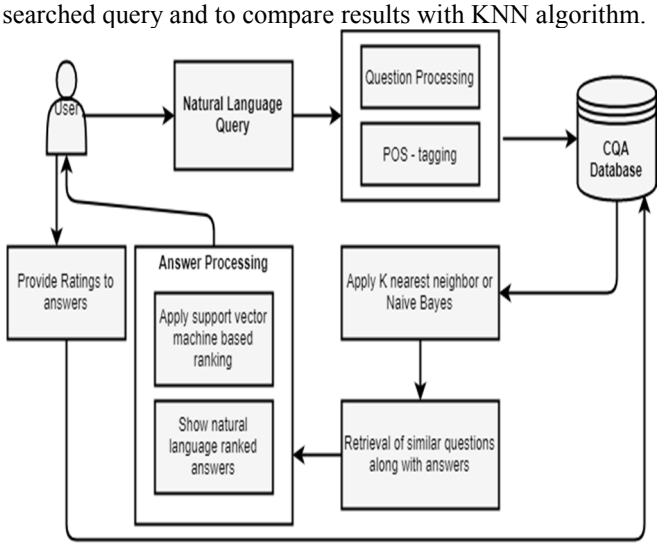

Fig.1 Proposed System Architecture

Rank Model: Rank model sorts all the answers related to the returned similar queries. It is already trained in offline learning and it takes question-answer pairs from online components which provide a ranking to pairs. It uses SVM based algorithm for finding the best relevant answers [1]. This method sorts the results using ranking methods based on how significant they are to the user query. It uses a mapping function to describe the match between a searched query and the features of each of the possible results returned.

B. Offline Module

This module randomly selects questions from the training set and establishes negative, positive and neutral training samples in the form of preferential pairs which are based on user ratings and votes. In this module, the rank model is trained based on preference pairs. The pairs are in the form of positive, negative and neutral pairs in terms of ratings or voting's given by users or asker of a question. Asker rates answer according to its relevance to question. These ratings are stored in datasets for future relevance findings.

C. Evaluation Method

For the experimental result, we have notations as follows:

TP: True Positive (number of cases which correctly retrieved),

FP: False positive (number of cases which incorrectly retrieved),

TN: True negative (correctly retrieved the number of cases as not required)

FN: False negative (incorrectly retrieved the number of cases as not required),

On the basis of this parameter, we can calculate two measurements

1. Precision: True Positives + False Positives != 0
   Precision = True Positives / (True Positives + False Positives + 0.1)

2. Recall: False Negatives = False Positive – True Positives;
   True Positive + False Negatives != 0
   Recall = True Positives / (True Positives + False Negatives + 0.1)

IV. Algorithms in Experiment

For the similar questions retrieval we have analyzed two algorithms K nearest neighbor and Naïve Bayes algorithm. After that output questions and answers are ranked using a support vector machine method.

A. KNN Algorithm

The K-Nearest Neighbor (KNN) is a machine learning algorithm whose goal is to classify objects into predefined classes of a sample group of documents. There is no need for training data to implement classification in KNN algorithm, this data can be used during the testing task. This algorithm is based on identifying the most relevant questions from sample groups of documents. For the retrieval of similar questions associated with the user query first question is processed and features are extracted.
Using TF-IDF method distance calculated with the query features. The Term frequency and inverse document frequency are the measurement methods which allows the calculation of the score for each word in every document. The technique finds the weight which evaluates the significance of terms in a collection of documents. The significance of the content is increased relatively to the number of occurring contents in the documents.

B. Naïve Bayes Algorithm

Naïve Bayes classifier is based on the probabilistic technique and depends on the Bayes theorem. In the supervised learning, the naïve Bayes classifier work. The particular attributes are considered.

Probability (a/b) = Probability (a/b)* Probability (a) / Prob. (b)

Probability (a/b) = rear probability

Probability (a) = preceding probability of class.

Probability (b/a) = potential probability of class

Probability (b) = preceding probability of predictor. On the basis of this algorithm, similar questions are retrieved.

C. Support Vector Machine Based Algorithm

In machine learning, support vector machine represents supervised learning methods with related learning algorithms. It determines data used for regression and classification analytics. SVM, Support Vector Machines helps to retrieve the best and most relevant questions from the input set of similar questions. Based on voting answers for the particular questions are ranked and most relevant questions with their answers are selected.

V. Implementation

A. Search Question

The user will search question on the following screen which will provide two options to search as k nearest neighbor search or naïve Bayes search.

After login into the system, the search question option will show where the user can search its question in the search box. Two options are provided to the user for comparison purpose. A query "what are symptoms of sinus infection?" is entered and after clicking on KNN search, the system will show above list of relevant questions to the user.

Ranked answers are shown when the question is expanded which is based on votes and ratings provided by users. After clicking on the rate the answer it shows the screen to rate answer according to low, medium and high.

When ratings are provided to the answer, rates are updated and remarks are changed for the answer.

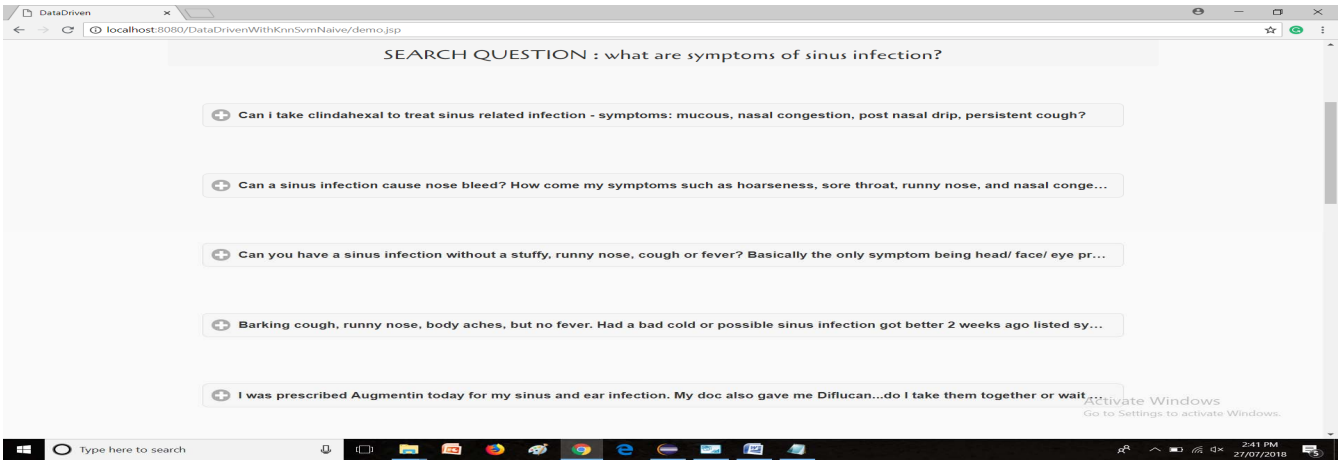

Fig.2 UI List of retrieved similar questions

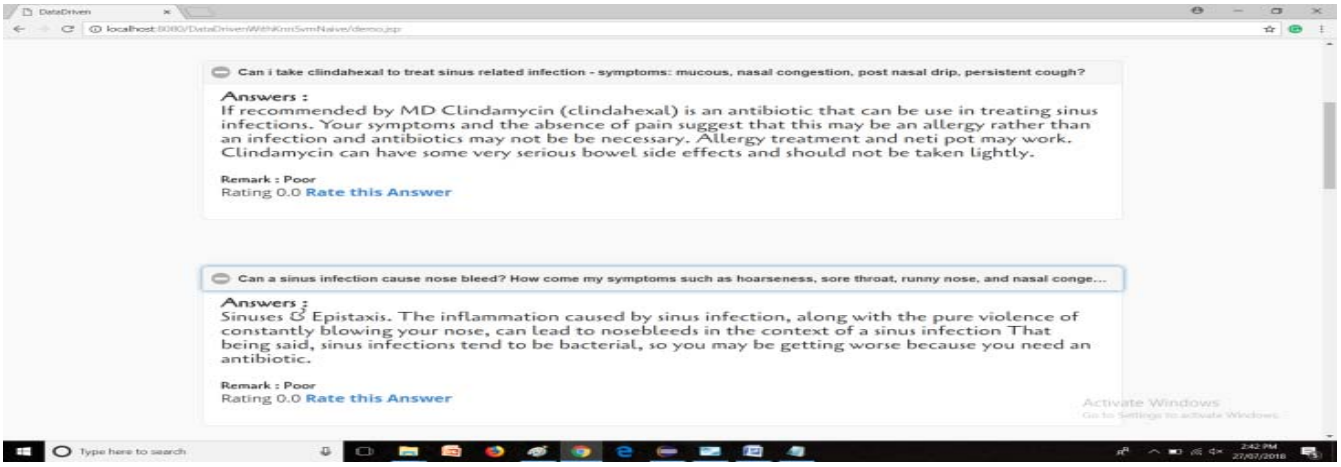

Fig.3 UI Question with ranked answer

VI. Results and Analysis

We have collected more than 35,000 questions and answers from websites to create QA dataset which contains multiple answers with respect to questions along with their ratings.

The following table 1 shows resulting values for user query "What are symptoms of blood cancer?" which represents K nearest neighbor returns less no. of relevant questions this gives more specific and precise results compared to naïve Bayes algorithm also time and memory required to process questions are more with this algorithm.

TABLE I. Comparison Chart For KNN and NB

Keywords	***Algorithms***	***Similar Question Count***	***Best Question Count***
Symptoms, Blood, Cancer	K nearest neighbor	5287	12
	Naïve Bayes	6961	29

Table 1.comparison of two algorithms in terms of no. of relevant questions returned according to the user's example query.

Also for the question "What are symptoms of sinus infection?" the keywords will be sinus, infection, symptoms etc. where for K nearest neighbor returns less no. of relevant questions this gives more specific and precise results than naïve Bayes algorithm.

The execution is assessed in terms of precision value and recall value where recall is a performance measure of the entire positive section of a dataset and precision is a performance measure of positive predictions. In results, it shows that K nearest neighbor along with SVM performs better compared to the Naïve Bayes algorithm.

Fig.4 shows precision and recall values for three different questions when applied K nearest neighbor algorithm for retrieving similar questions and SVM for retrieving best question along with answers whereas Fig. 5 shows precision and recall values for three different questions when applied Naïve Bayes algorithm for retrieving similar questions and SVM for retrieving best question along with answers. The graph shows K nearest neighbor gives better performance compared to the Naïve Bayes algorithm.

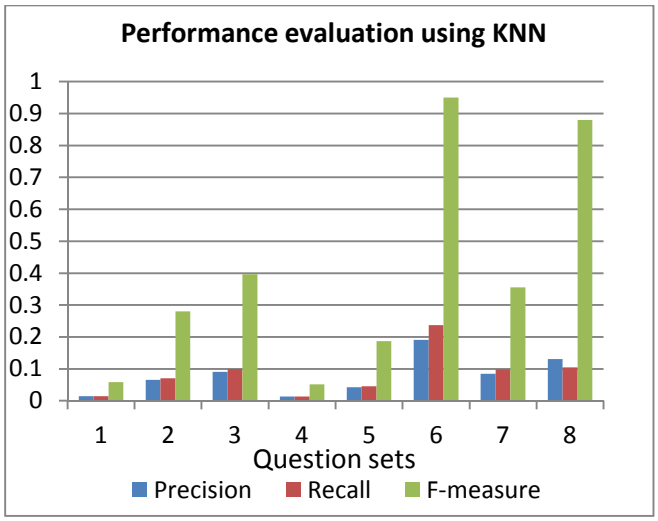

Fig.4 Performance analysis (precision, recall and f-measure) using KNN algorithm

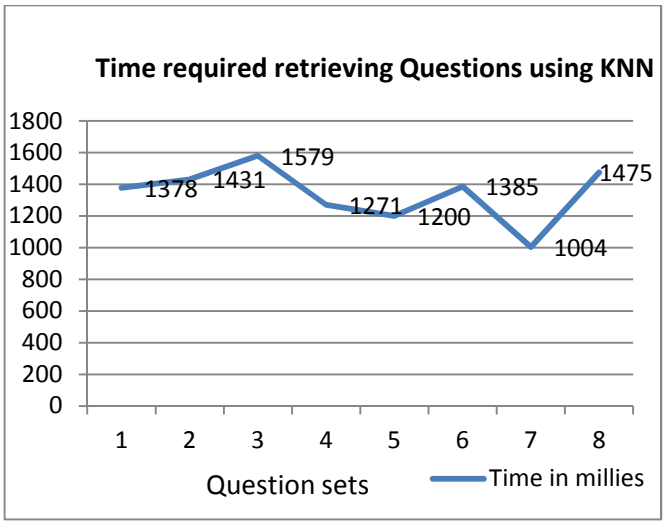

Fig.6 Time required retrieving Questions using K nearest neighbor algorithm.

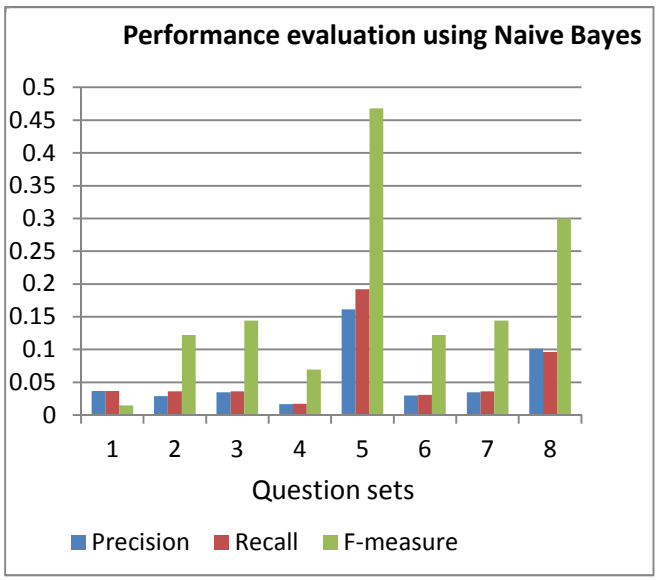

Fig.5 Performance Analysis (precision, recall and f-measure) using Naïve Bayes algorithm.

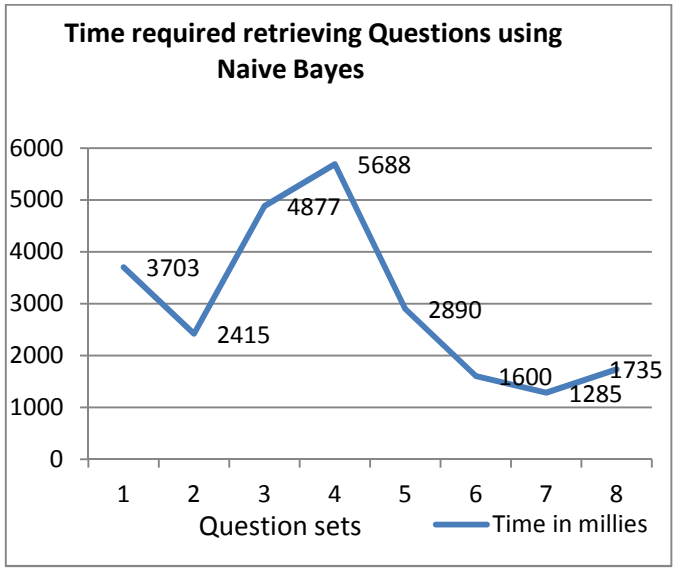

Fig.7 Time required retrieving Questions using Naïve Bayes algorithm.

The time graph represents the time required to retrieve a number of questions using different algorithms with respect to system count.

According to graphs as shown in figure time required to process and retrieve questions along with answers using K nearest neighbor is less compared to naïve Bayes algorithm. Time is calculated in terms of Millis against the system running count.

Also, no. of similar questions retrieved with knn are less compared to naïve Bayes which proportionally requires less time and memory. Following graph Fig. 8 shows a count of similar questions retrieved using K nearest neighbor algorithm where from 35,000 questions knn retrieves most similar questions in fewer numbers and support vector machine selects most similar and relevant questions to searched question.

From graph in figure 8, the count of similar questions retrieved using Naïve Bayes algorithm where from 35,000 questions NB retrieves most similar questions in more numbers than knn and support vector machine selects most similar and relevant questions to searched question as it requires more time and memory.

VII. CONCLUSION

The work includes a way to find the best and relevant answers to asked questions from previously asked similar questions where two different algorithms Knn and Naïve Bayes are compared. We got better results for Knn compared to Naïve Bayes with results as well as in terms of time required. In this system, first, similar questions are retrieved for a given question and pool of answers are collected which are given to rank model where it ranks answers based on features which will the provide user most relevant answers in less time, also user can rate the hose answers for further retrieval.

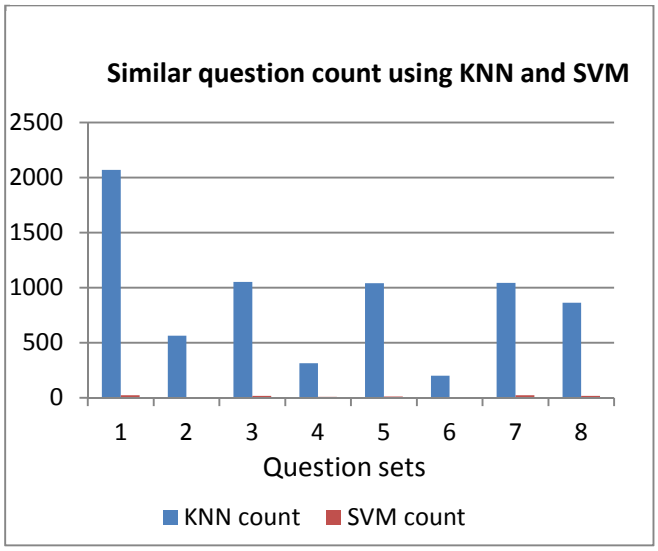

Fig. 8 Similar question count using k nearest neighbor algorithm.

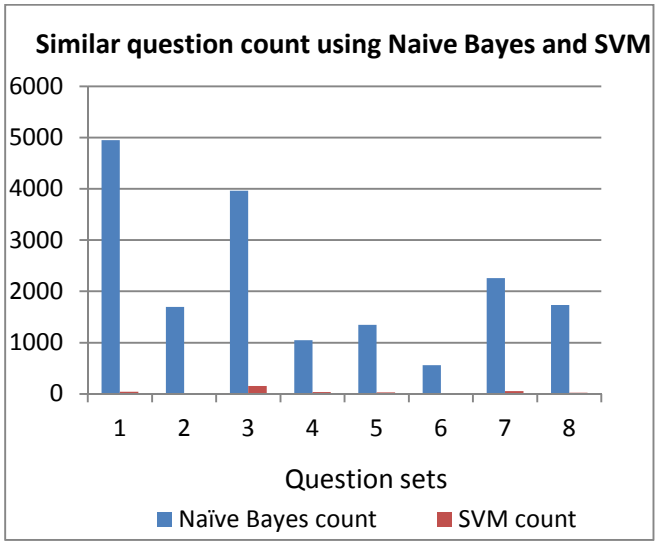

Fig. 10 Similar question count using a Naïve Bayes algorithm.

In future noisy data can be handled related to multi-topical questions for that work can be done related to topic modeling for general community question-answering to improve answer ranking. Also, Query-focused summarization can be implemented to provide a summary to a new question from historic archives where the user will get a summary of multiple answers so that one complete and the most relevant answer can be formed.

REFERENCES

[1] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, "Datadriven Answer Selection in Community QA Systems", *IEEE transactions on knowledge and data engineering*, June 2016

[2] Z. Ji and B. Wang, "Learning to rank for question routing in community question answering," *in Proceedings of CIKM'13.ACM*, 2013, pp. 2363–2368.

[3] Y. Tian, P. S. Kochhar, E.-P. Lim, F. Zhu, and D. Lo, ''Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting,'' *in Proc. Workshops Int. Conf. Social Informat.*, 2013, pp. 55–68.

[4] Sahu, T. P., Nagwani, N. K., & Verma, S. "Selecting Best Answer: An Empirical Analysis on Community Question Answering Sites", *IEEE Access*, 4, 4797–4808. doi:10.1109/access.2016.

[5] Calefato, F., Lanubile, F., & Novielli, N." Moving to stack overflow: Best-answer prediction in legacy developer forums". *In Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement. Article* 13(pp. 1–10). ACM,2016

[6] F. Hieber and S. Riezler, "Improved answer ranking in social question-answering portals," *in Proceedings of SMUC'11. ACM*, 2011, pp. 19–26

[7] Liu Yang,Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang," CQARank: Jointly Model Topics and Expertise in Community Question Answering", *Institutional Knowledge at Singapore Management University*,2013

[8] Daniel Hasan Dalip, Marco Cristo,Pável Calado," Exploiting User Feedback to Learn to Rank Answers in Q&A Forums: a Case Study with Stack Overflow",*ACM*,2013

[9] Pradeep Kumar Roy, Zishan Ahmad, Jyoti Prakash Singh ,"Finding and Ranking High-Quality Answers in Community Question Answering Sites", *Global Journal of Flexible Systems Management Springer*,November 2017

[10] Guoxin Liu, Haiying Shen," iASK: A Distributed Q&A System Incorporating Social Community and Global Collective Intelligence", *IEEE transaction* 2017

[11] Dalia Elalfy, Walaa Gad, Rasha Ismail, "Predicting Best Answer in Community Questions based on Content and Sentiment Analysis", *IC1CIS*'15

[12] Oleksandr Kolomiyets, Marie-Francine Moens,"A Survey on Question Answering Technology from an Information Retrieval Perspective", *publication in Information Sciences*, August 2011

[13] W. Wei, Z. Ming, L. Nie, G. Li, J. Li, F. Zhu, T. Shang, and C. Luo, "Exploring heterogeneous features for query-focused summarization of categorized community answers*," Inf. Sci.,* vol. 330, pp. 403–423, 2016.

[14] Zongcheng Ji, and Bin Wang,"Learning to Rank for Question Routing in Community Question Answering",*ACM* 2013

[15] Show-Jane Yen, Yu-Chieh Wu, Jie-Chi Yang, Yue-Shi Lee,"A support vector machine-based context-ranking model for question answering", *Information Sciences*,october 2012

[16] QuanHungTran,VuDucTran,TuThanhVu,MinhLeNguyen,SonBaoPham,"JAIST:CombiningmultiplefeaturesforAnswerSelectioninCommunity QuestionAnswering", *SemEval* 2015

[17] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan ,"Disease Inference from Health-Related Questions via Sparse Deep Learning", *IEEE transactions on knowledge and data engineering*, May 2014

[18] Zhou Zhao, Hanqing Lu, Deng Cai, Xiaofei He, and Yueting Zhuang. "User preference learning for online social recommendation" *,IEEE Trans. Knowl. Data Eng., 28(9):2522–2534*, 2016.

[19] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. "Expert finding for community-based question answering via ranking metric network learning". *In IJCAI pages 3000–3006*, 2016.

[20] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. "Question answer topic model for question retrieval in community question answering". *In CIKM, pages 2471–2474. ACM*,2012.