Finding High-Quality Content in Social Media

Group - 16
Pankaj Kukreja (CS15BTECH11029)
Nidhi Dhamnani (CS15BTECH11028)
Sahil Yerawar (CS15BTECH11044)
Sambit Rath (ES15BTECH11015)

Outline

- Introduction
- Related Work
- Problem Approach
- Experiments and Results
- Week-1 Implementation

Introduction

- The quality of user-generated content varies drastically from excellent to abuse and spam.
- Social media in general exhibit a rich variety of information sources: in addition to the content itself, there is a wide array of non-content information available, such as links between items and explicit quality ratings from members of the community.
- As the availability of such content increases, the task of identifying high-quality content in sites based on user contributions—social media sites becomes increasingly important.

Introduction (Contd.)

- User-generated content has become increasingly popular on the web, more and more users participate in content creation, rather than just consumption.
- An important difference between user-generated content and traditional content that is particularly significant for knowledge-based media such as question/answering portals is the variance in the quality of the content.
- The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive content.
- Social media exhibit a wide variety of user-to-document relation types, and user-to-user interactions, which is modelled as graph-based framework.

Source of Data Collection

- The paper focuses on **Yahoo! Answers**.
- Yahoo! Answers is a question/answering system where people ask and answer questions on any topic.
- Community question/answering portals are a popular destination of users looking for help with a particular situation, for entertainment, and for community interaction.
- A user can vote for answers of other users, mark interesting questions, and even report abusive behavior.
- The central element of the Yahoo! Answers system are questions. Each question has a life cycle. It starts in an “open” state where it receives answers.

Source of Data Collection (Contd.)

- Then at some point , the question is considered “closed,” and can receive no further answers. At this stage, a “best answer” is selected either by the asker or through a voting procedure from other users; once a best answer is chosen, the question is “resolved.”
- The system is partially moderated by the community,
  - Any user may report another user’s question or answer as violating the community guidelines (e.g., containing spam, adult-oriented content, copyrighted material, etc.)
  - A user can also award a question a “star”, marking it as an interesting question.
  - A user can sometimes vote for the best answer for a question, and can give to any answer a “thumbs up” or “thumbs down” rating, corresponding to a positive or negative vote respectively.

Related Work

- Link analysis in social media: Link-based ranking algorithms were successful in estimating the quality of web pages and have been applied in this context. Two of the most prominent link-based ranking algorithms are PageRank and HITS.
- Propagating reputation: It involves the study of propagating trust and distrust among users. Trust can be considered as a transitive property whereas distrust cannot be considered transitive property. They present a taxonomy of trust metrics and discuss ways of incorporating information about distrust into the rating scores.
- Question/answering portals and forums: According to a study, the quality of answers in question/answering portals is good on average, but the quality of

Related Work (Contd.)

specific answers varies significantly. In particular, in a study of the answers to a set of questions in Yahoo! Answers, the authors found that, the fraction of

a. Correct answers to specific questions = 17%-45%
b. At least one good answer = 65%-90%, which means a method for finding high-quality answers can have a significant impact in the user's satisfaction with the system.

The paper focuses on features derived from the particular answer being analyzed, such as answer length, number of points received, etc., as well as user features, such as fraction of best answers, number of answers given, and by identifying quality of questions in addition to answer quality.

Related Work (Contd.)

- Expert finding: There is a high correlation between link-based metrics and the answer quality. According to a study, HITS on user-answer graph is a promising approach, as the obtained authority score is better correlated with the number of votes that the items receive, than simply counting the number of answers the answerer has given in the past. In real data, ExpertiseRank outperforms HITS.
- Text analysis for content quality: In Automated Essay Grading (AES) writings of students are graded by machines. AES are built as text classification tools, and use a range of properties derived from the text as features. A different area of study involving text quality is readability. The approach is to combine the number of syllables or words in the text with the number of sentences—the first being a

Related Work (Contd.)

crude approximation of the syntactic complexity and the second of the semantic complexity.

- Implicit feedback for ranking: Implicit feedback from web users has been shown to be a valuable source of result quality and ranking information. The results of *click interpretation* are applied as a source of quality information in social media.

Content Quality Analysis in social media

- The approach is to exploit features of social media that are intuitively correlated with quality, and then train a classifier to appropriately select and weight the features for each specific type of item, task, and quality definition.
- The following feature types are used as an input to a classifier that can be tuned for the quality definition for the particular media type:
  - Intrinsic Content Quality
  - User Relationships
  - Usage Statistics
- The problem of quality ranking is casted as a binary classification problem. Experiments were done on various algorithms such as SVMs, Decision Trees and Gradient Boosted trees (best performing algorithm).

Content Quality Analysis in social media (Contd.)

- **Intrinsic content quality**: The intrinsic quality metrics (i.e., the quality of the content of each item) used are mostly text-related as the social media items evaluated are primarily textual in nature. n-grams up to length 5 that appear in the collection more than 3 times are used as features. The following semantic features are used:
  - Punctuation and typos: Common ill practices are ignored
  - Syntactic and semantic complexity: Simple proxies for complexity are simplified.
  - Grammaticality: Linguistic features are used such as part of speech, n-grams
- **User relationships**: A significant amount of quality information can be inferred from the relationships between users and items. Link-analysis algorithms can be used for propagating quality scores in the entities of the question/answer. These

Content Quality Analysis in social media (Contd.)

relationships are represented as edges in a graph, with content items and users as nodes. The edges are typed, i.e., labeled with the particular type of interaction. From these graphs the hubs and authorities scores and the **PageRank scores** are computed.

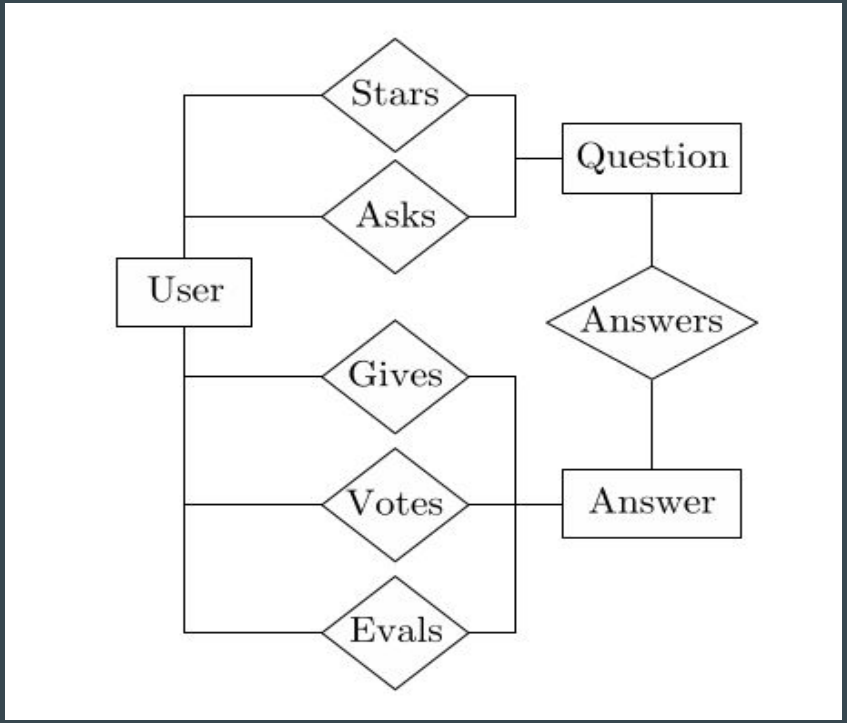
- **Usage statistics**: Usage statistics such as the number of clicks on the item and dwell time have been shown useful in the context of identifying high quality web search results, and are complementary to link-analysis based methods.

Modeling Content Quality in Community Question/Answer

- Yahoo! Answers is question-centric i.e the interactions of users are organized around questions.
- The main forms of interaction among the users are
  - Asking Questions
  - Answering a Question
  - Selecting Best Answer
  - Voting an Answer

Application-specific user relationships

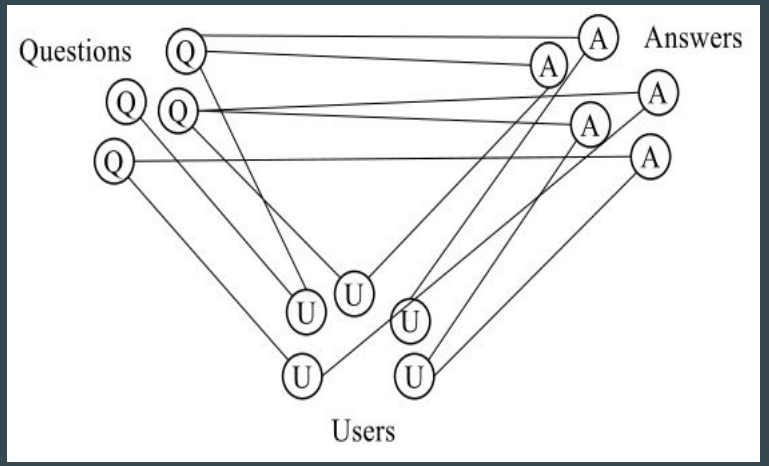
- The relationship between user, answer and question can be shown using entity relationship model as shown in the figure.



Entity-relationship diagram

Application-specific user relationships

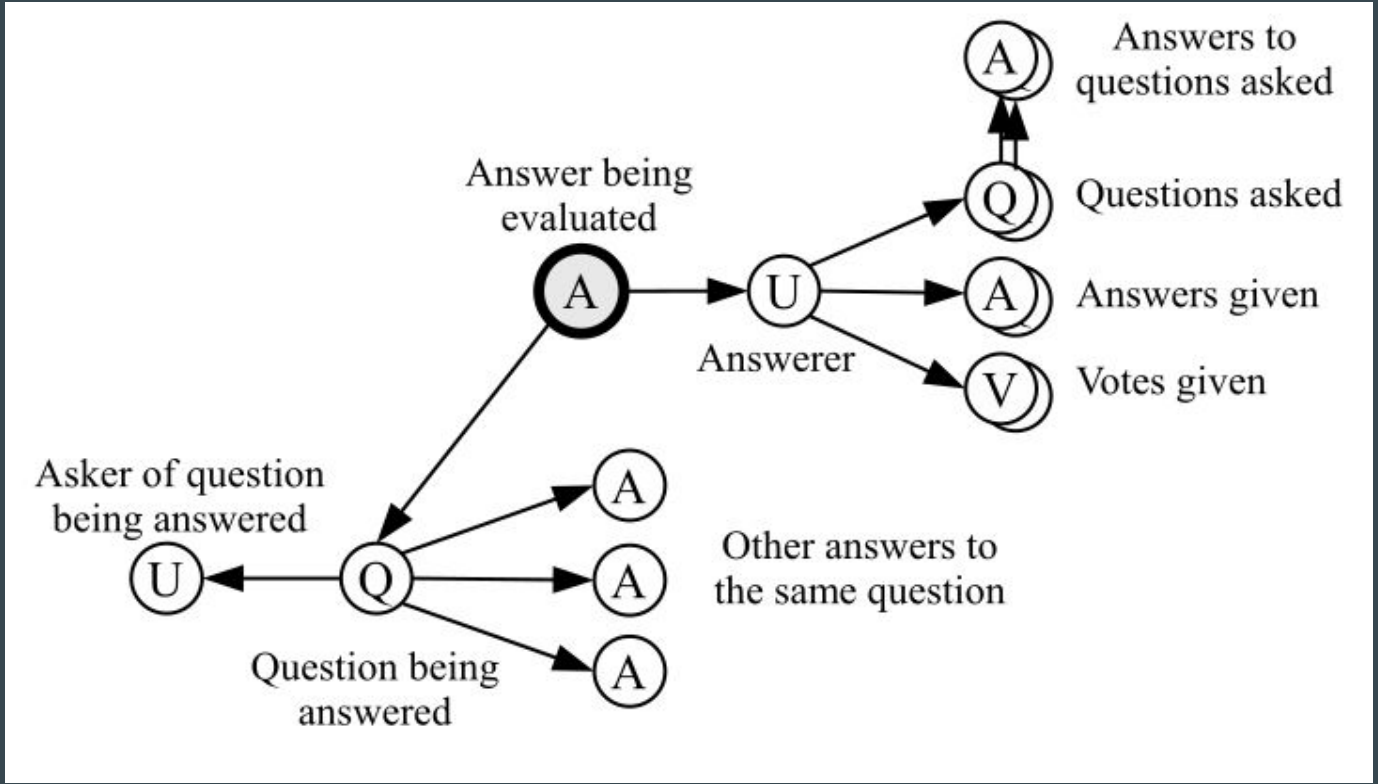
- The relationships between questions, users asking and answering questions, and answers can be captured by a tripartite graph where an edge represents an explicit relationship between the different node types.
- Since a user is not allowed to answer his/her own questions, there are no triangles in the graph, so in fact all cycles in the graph have length at least 6.



Interaction of users-questions-answers modeled as a tripartite graph

Application-specific user relationships (Contd.)

- We use multi-relational features to describe multiple classes of objects and multiple types of relationships between these objects.
- Each feature is characterized by the path from the root of the tree to that node. Hence, each specific feature can be represented by a path in the tree (following the direction of the edges).
- For instance, a feature of the type “QU” represents the information about a question (Q) and the user (U) who asked that question.



Types of features available for inferring the quality of an answer

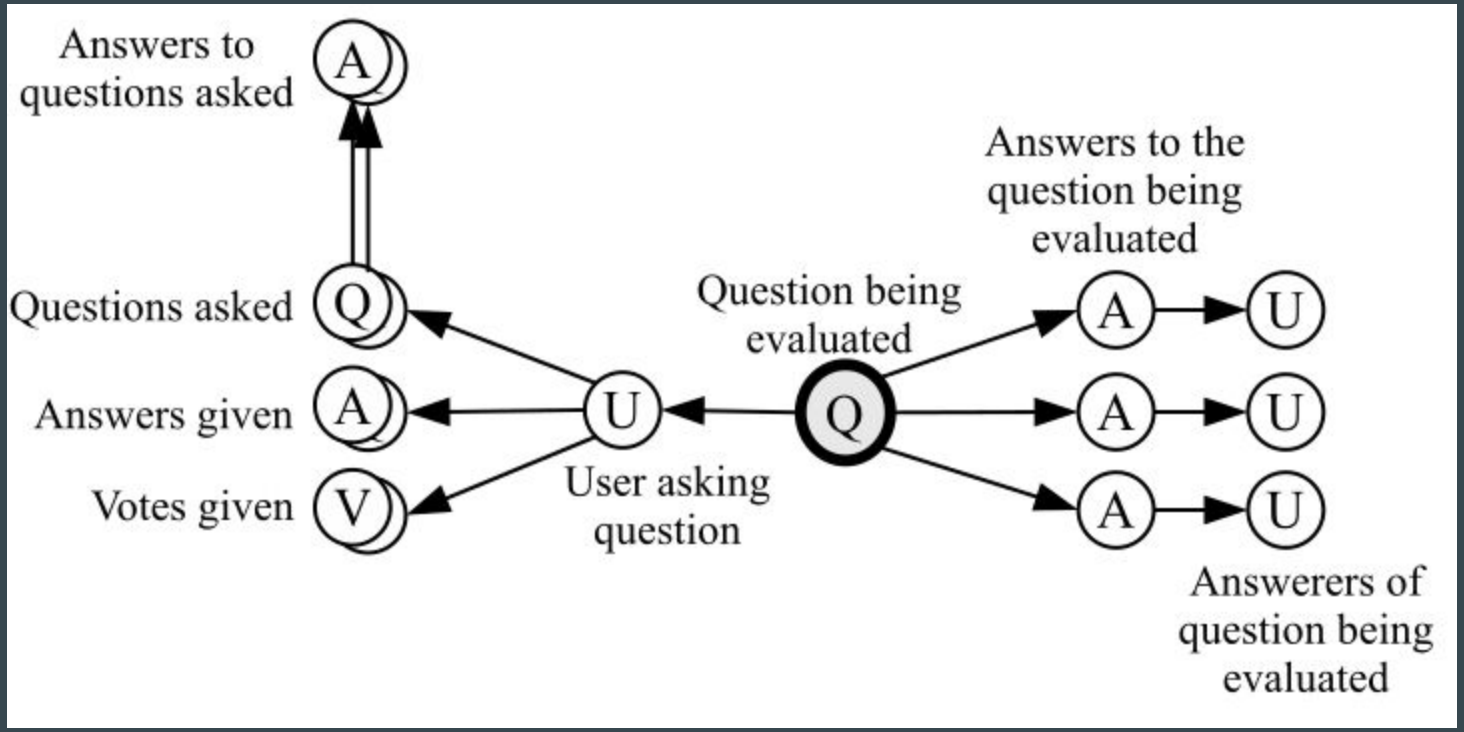
Features for inferring the quality of an answer

The types of features on the **question subtree** are:

- Q Features from the question being answered
- QU Features from the asker of the question being answered
- QA Features from the other answers to the same question

The types of features on the **user subtree** are:

- UA Features from the answers of the user
- UQ Features from the questions of the user
- UV Features from the votes of the user
- UQA Features from answers received to the user's questions U Other user-based features



Types of features available for inferring the quality of a question

Features for inferring the quality of an Question

- There are two subtrees: one related to the asker of the question, and the other related to the answers received
- The types of features on the answers subtree are:
  - “A” Features directly from the answers received
  - “AU” Features from the answerers of the question being answered
- The types of features on the user subtree are the same as the ones above for evaluating answer.

Implicit user-user relations

We consider the user-user graph $G = (V, E)$ in which set of vertices V is composed of the set of users and the set $E = E_a \cup E_b \cup E_v \cup E_s \cup E_+ \cup E_-$ represents the relationships between users as follows:

- E_a represents the answers: $(u, v) \in E_a$ iff user u has answered at least one question asked by user v.
- E_b represents the best answers: $(u, v) \in E_b$ iff user u has provided at least one best answer to a question asked by user v.
- E_v represents the votes for best answer: $(u, v) \in E_v$ iff user u has voted for best answer at least one answer given by user v.
- E_s represents the stars given to questions: $(u, v) \in E_v$ iff user u has given a star to at least one question asked by user v.
- E_+ / E_- represents the thumbs up/down: $(u, v) \in E_+ / E_-$ iff user u has given a “thumbs up/down” to an answer by user v.

Implicit user-user relations

- For each graph G x = (V, E_x), we denote by
  - h_x the vector of hub scores on the vertices V
  - a_x the vector of authority scores
  - p_x the vector of PageRank scores
  - p'_x the vector of PageRank scores in the transposed graph.
- To classify these features in our framework, we consider that PageRank and authority scores are related mostly to in-links, while the hub score deals mostly with out-links. For instance, let's take h_b . It is the hub score in the "best answer" graph, in which an out-link from u to v means that u gave a best answer to user v. Then, h_b represents the answers of users, and is assigned to the answerer record (UA).

Content Feature for QA

- We rely on feature selection methods and the classifier to identify the most salient features for the specific tasks of question or answer quality classification.
- Additionally, we devise a set of features specific to the QA domain that model the relationship between a question and an answer.
    - A copy of a Wall Street Journal article about economy may have good quality, but would not (usually) be a good answer to a question about celebrity fashion.
- We model the relationship between the question and the answer using the KL-divergence (relative entropy) between the language models of the two texts, their non-stopword overlap, the ratio between their lengths, and other similar features.

Usage features for QA

- A question thread is usually viewed as a whole, and the content usage statistics are available primarily for the complete question thread. As a base set of content usage features we use the number of item views.
- We also use the metadata available for each question
  - Duration for which question was open
  - Average view count (or click count) in the genre in which question was posted
- We also normalize the value of features
  - Click frequency normalized by subtracting the expected click frequency for that category, divided by the standard deviation of click frequency for the category.

Experimental Setup

- For experimental setting we take a dataset which consists of over 6000 questions and over 8000 question & answer pairs.
- All of the questions and answers were labeled by human editors.
- Questions and answers were graded based on how well-formed they were, how readable they were and also based on their utility.

Experimental Setup (Contd.)

- We build evaluation dataset from the data we collected earlier. We separate questions and answers into two sets. Let Q_0 be the set of questions included in evaluation dataset and A_0 be the set of answers included in evaluation dataset.
- Let U_1 be the set of users who either made a question which belongs to Q_0 or gave an answer which belongs to A_0 and let Q_1 be the set of all questions asked by users who belong to U_1.
- Clearly, $Q_0 \subseteq Q_1$
- Similarly we select A_1 to be the set of answers given by users in U_1 and A_2 to be the set of all the answers to questions in Q_1
- As we can see clearly $A_0 \subseteq A_1$. We define our dataset by the nodes (Q_1, $A_1 \cup A_2$, U_1)

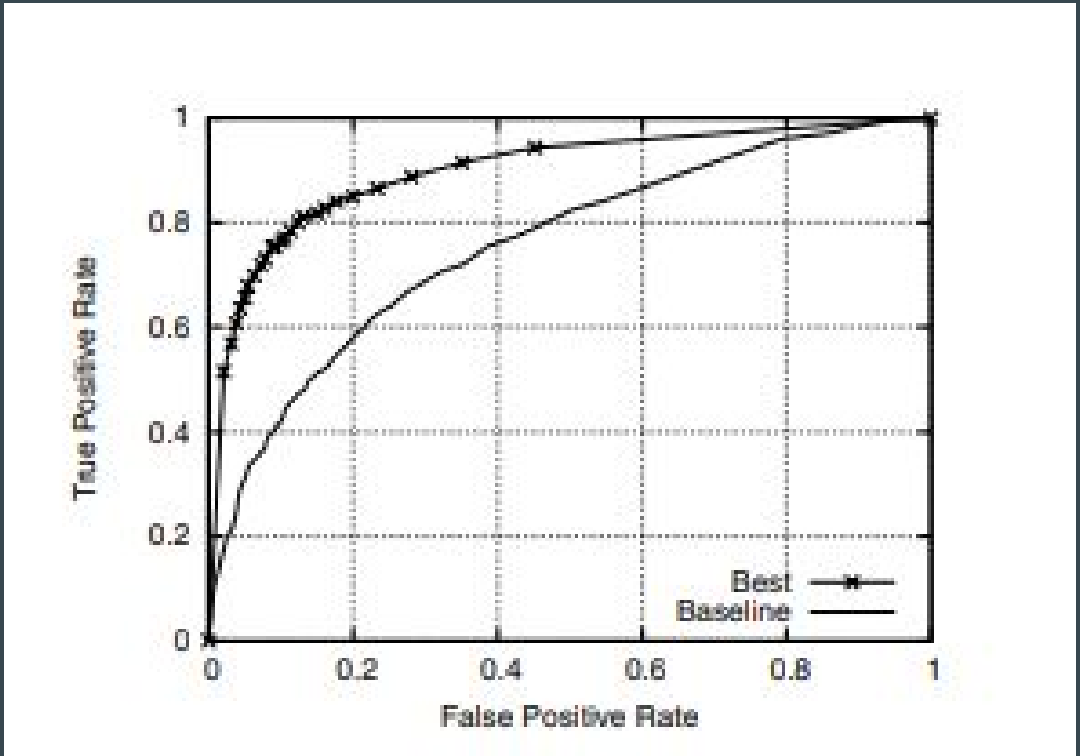
Experimental Setup (Contd.)

- We define G_x as the graph with vertices V and Edges E_x where $x \in \{a, b, v, s, +, -\}$
- We computed the hubs ,authorities scores and pagerank scores as we did in HITS algorithm.
- After doing all these experimental setup we see the results obtained and show them in form of graphs and tables.

Experimental Results

- Question Quality
  - Approaches taken: Text(n-gram model), Content Based Features, Usage Based and Relational
  - Results show gradual increase in performance when additional information is used.
  - For Content Based Usage features, topic information is important
  - Chances of Overfitting when relying on one metric alone.
  - Some Significant features: Category, Normalized clickthrough

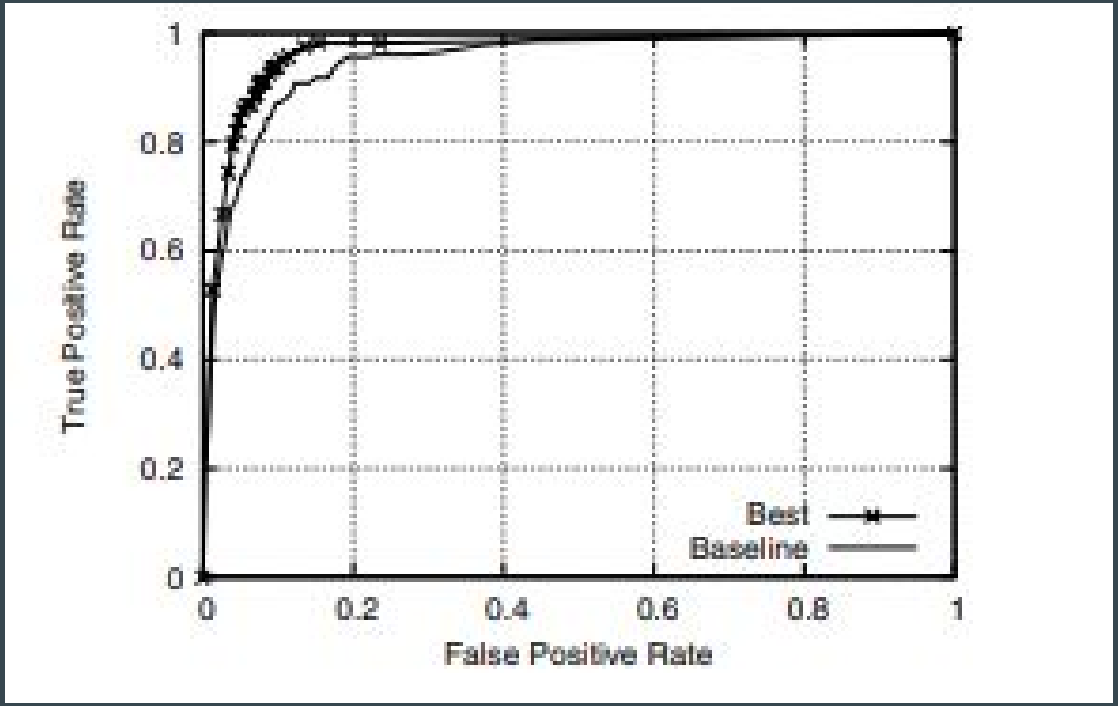
Question Quality Graph



Experimental Results

- Answer Quality
  - Same evaluation methodology as before, only usage count is not used here.
  - High scores indicate system performance closer to that of humans.
  - Multiple feature sets show improved performance, though substantially
  - Predicted features for answer quality have non-uniform weight, where few of them have unfair advantage over other.

Answer Quality Graph



Conclusion

- General Framework for quality estimation
- Graph-based modelling along with multiple features sets considered before provide high-levels of estimates on Q/A quality.
- Due to low levels of dependence among these feature sets, it can increase classifier's robustness to spam.
- Work can be extended to use the same methodology to identify malicious users in community Q/A portals.

Week-1: Implementation

- Checked the performance of various algorithms on email dataset which contains non spam (high quality content) and spam (low quality content) mails.
  - Algorithms used (Reference from paper):
    - Naive Bayes
    - Decision Trees
    - Gradient Boosting
  - Best Performing Algorithm: **Gradient Boosting**!
- Requested for Yahoo Q&A Dataset.

Additional Reference

- http://www.cs.cmu.edu/~ark/QA-data/
- https://github.com/bdhingra/quasar
- https://www.kaggle.com/stanfordu/stanford-question-answering-dataset/version/1
- https://www.researchgate.net/post/What_are_the_datasets_available_for_question_answering_system
- https://github.com/karthikncode/nlp-datasets#question-answering

Thank You!

Implicit user-user relations

The assignment of these features is done in the following way:

- UQ To the asker record of a user: a_a , a_b , a_s , p_a , p_b
- UA To the answerer record of a user: h_a , h_b , p_a , p_b , a_v , p_v , a_+ , p_+ , a_- , p_-
- UV To the voter record of a user: h_v , p_v , h_s , p_v , h_+ , p_+ , h_- , p_-