Identifying Textual Features of High-Quality Questions:

An Empirical Study on Stack Overflow

Qing Mi[†], Yujin Gao[‡], Jacky Keung[†], Yan Xiao[†], Solomon Mensah[†]

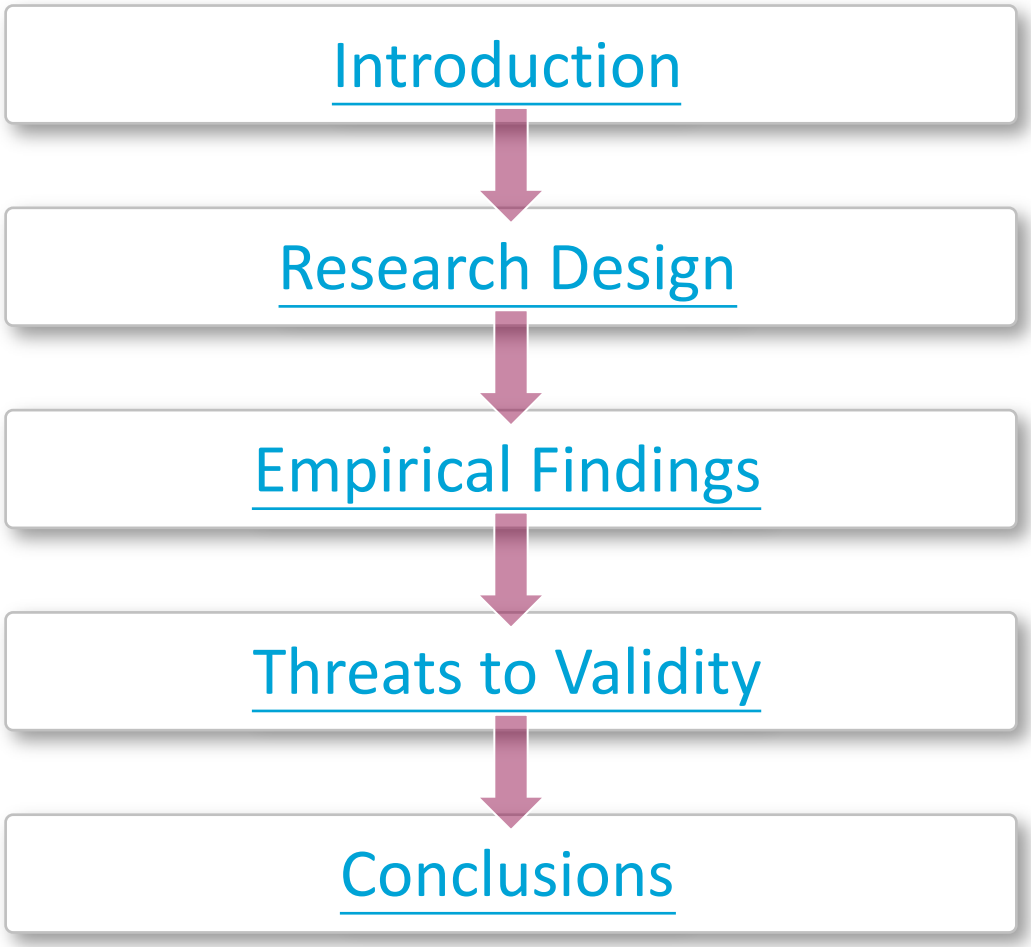
[†]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

[‡]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

CityU 香港城市大學 City University of Hong Kong

Department of Computer Science

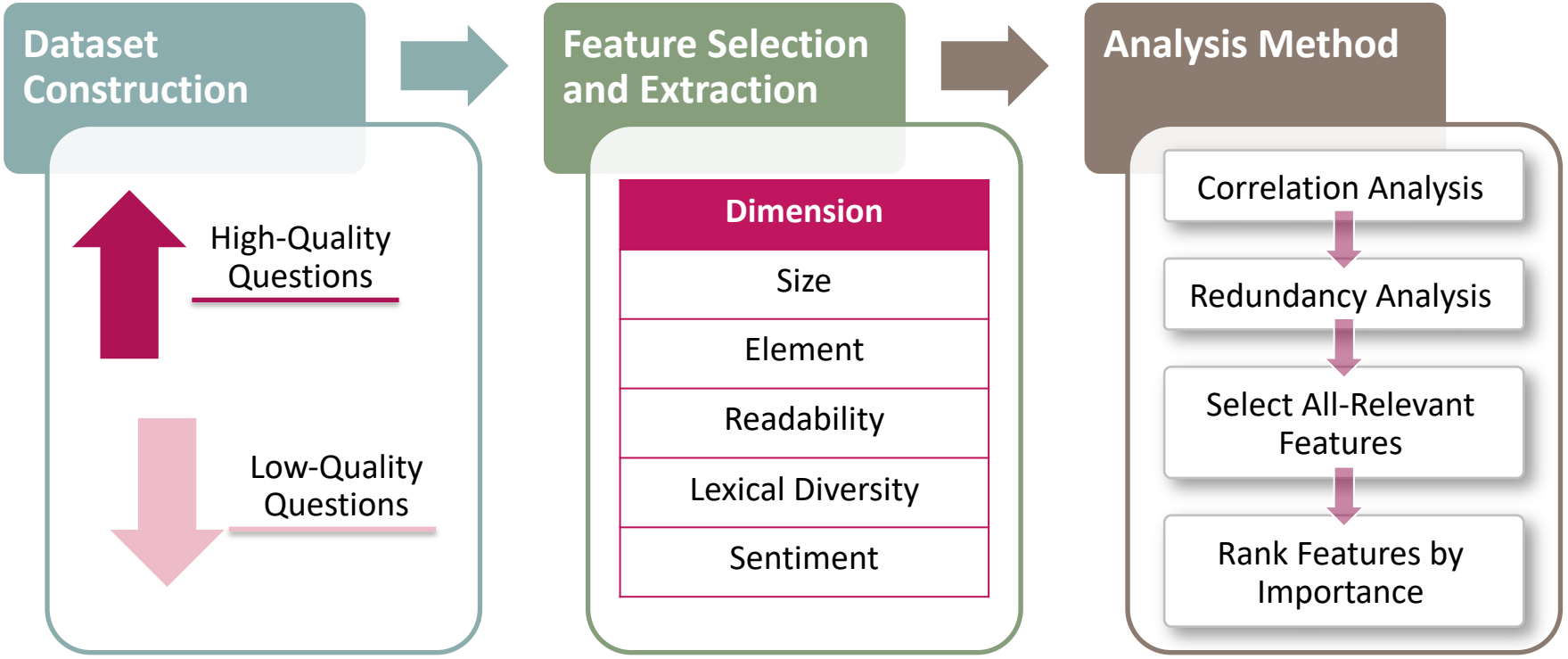
Presentation Outline



Context and Problem Statement

- Stack Overflow
  - A programming-specific Q&A website.
  - A valuable repository of software engineering knowledge.
- Research Objective
  - Identify textual features of high-quality questions on Q&A websites.

Research Design



Dataset Construction

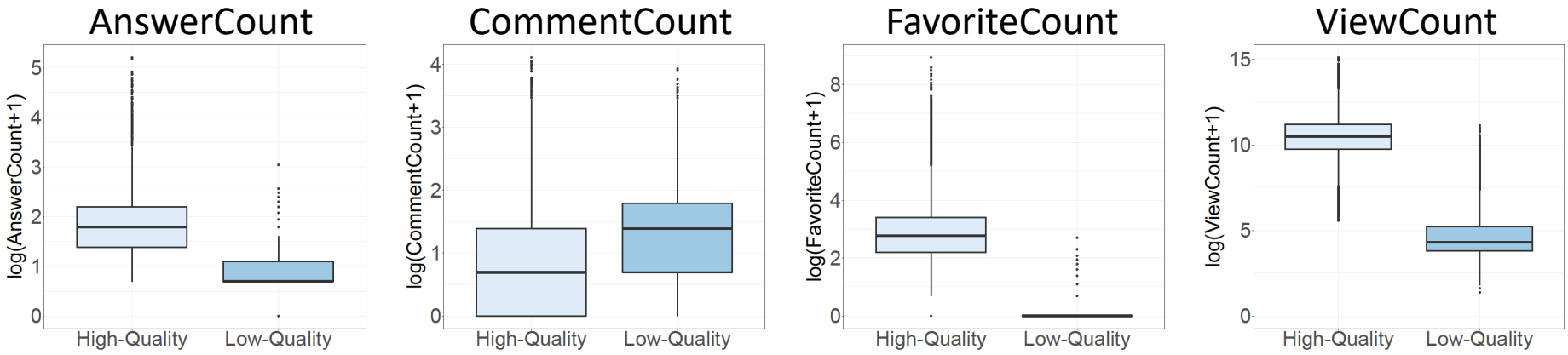
- Raw Dataset
  - Stack Overflow Data Dump: XML-formatted files.
  - Posts.xml: 12.35M questions and 19.78M answers.
- Question Quality
  - Voting Score: the difference between upvotes and downvotes on the user post.

Dataset Construction

- **High-Quality Group**
  The top 1% questions with accepted answers.
  Average voting score: 84.84

vs

- **Low-Quality Group**
  The bottom 1% questions without accepted answers.
  Average voting score: -3.96



Boxplots (on Logarithmic Scale) Comparing High-Quality Questions with Low-Quality Questions

Feature Selection and Extraction

- Why Textual Features?
  - They can be easily improved by questioners with appropriate actions.
  - The findings can be generalized to other Q&A websites.
- Which Textual Features?
  - Size
  - Element
  - Readability
  - Lexical Diversity
  - Sentiment

Feature Selection and Extraction

- Size
  - Combine the title and the content of each question.
  - Remove all non-text elements (e.g., images).

Textual Features Considered in This Study

Dimension	Feature	Description
Size	Paragraphs	The number of paragraphs in the question.
	Sentences	The number of sentences in the question.
	Words	The number of words in the question.
	Letters	The number of letters in the question.
	AvgParaLen	The average number of sentences per paragraph.
	AvgSntcLen	The average number of words per sentence.
	AvgWordLen	The average number of letters per word.

Feature Selection and Extraction

- Element
  - Count the number of the selected elements using HTML tags: <ul> and <ol> for Lists, and for EmphTexts.

Textual Features Considered in This Study

Dimension	Feature	Description
Element	Lists	The number of ordered/unordered lists in the question.
	EmphTexts	The number of emphasized/strong texts in the question.
	Links	The number of hyperlinks in the question.
	CodeSnippets	The number of code snippets in the question.
	Tags	The number of tags that describe the topic of the question.

Feature Selection and Extraction

- Readability
  - A human judgment on how easy it is to understand a text.
- Reading Grade Level
  - Easy for <6
  - Average for 7-9
  - Difficult for >9

Feature Selection and Extraction

Textual Features Considered in This Study

Dimension	Feature	Description
Readability	ARI	Automated Readability Index $= 0.5\times\frac{W}{S_t} + 4.71\times\frac{C}{W} - 21.43$
	SMOG	Simple Measure of Gobbledygook $= 1.043\times\sqrt{W_{3S_y}\times\frac{30}{S_t}} + 3.1291$
	Flesch	Flesch Reading Ease $= 206.835 - 1.015\times\frac{W}{S_t} - 84.6\times\frac{S_y}{W}$
	GunningFog	Gunning Frequency of Gobbledygook $= 0.4\times\left(\frac{W}{S_t} + \frac{100\times W_{3S_y}}{W}\right)$
	FleschKincaid	Flesch Kincaid Grade Level $= 0.39\times\frac{W}{S_t} + 11.8\times\frac{S_y}{W} - 15.59$
	FORCAST	FORCAST $= 20 - \frac{W_{1S_y}\times 150/W}{10}$
	ColemanLiau	Coleman Liau $= 5.88\times\frac{C}{W} - 29.6\times\frac{W}{S_t} - 15.8$

S_t stands for the number of sentences, W for the number of words, C for the number of characters, S_y for the number of syllables, W_{1Sy} for the number of words with exactly one syllable, W_{3Sy} for the number of words with at least three syllables.

Feature Selection and Extraction

❖ Lexical Diversity

- The range of different word stems used in a text.
- An important measurement of text difficulty.

Textual Features Considered in This Study

Dimension	Feature	Description
Lexical Diversity	Maas	The Maas Index $= \frac{\log \mathrm{N} - \log V}{\log N \times \log N}$
	MTLD	The average number of sequential words in a text that maintain a certain TTR value.
	HDD	For each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words.

N stands for the number of tokens, V for the number of types, TTR for the classic type-token ratio.

Feature Selection and Extraction

❖ Sentiment

- The sentence-level result: a dual tuple (P, N).

  P: 1 (not positive) to 5 (extremely positive)
  N: -1 (not negative) to -5 (extremely negative)

- The document-level result: the polarity of sentiment.

  Positive for $M_P + M_N > 0$
  Negative for $M_P + M_N < 0$
  Neutral for $M_P + M_N = 0$

Textual Features Considered in This Study

Dimension	Feature	Description
Sentiment	PosScore	The maximum positive sentiment strength.
	NegScore	The maximum negative sentiment strength.
	SentiScore	The document-level sentiment strength.

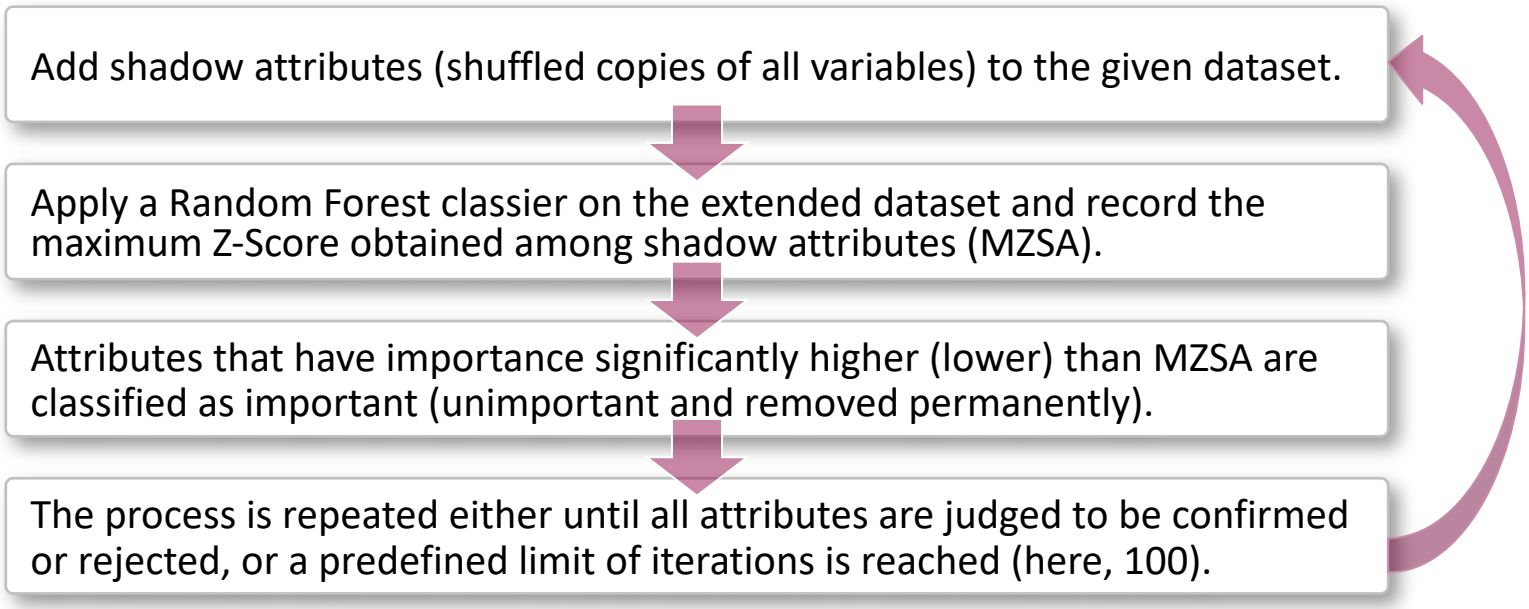
Analysis Method

- Step 1: Correlation Analysis
  - To detect collinearity between features, we perform a variable clustering analysis.
  - For the highly correlated variables, we reserve only one from each pair.
- Step 2: Redundancy Analysis
  - To remove redundant features, we examine to what extent each variable can be predicted from the remaining ones.
  - At each step, the most predictable variable is dropped.

Analysis Method

❖ Step 3: Select All-Relevant Features

The Boruta Algorithm (An All-Relevant Feature Selection Method)

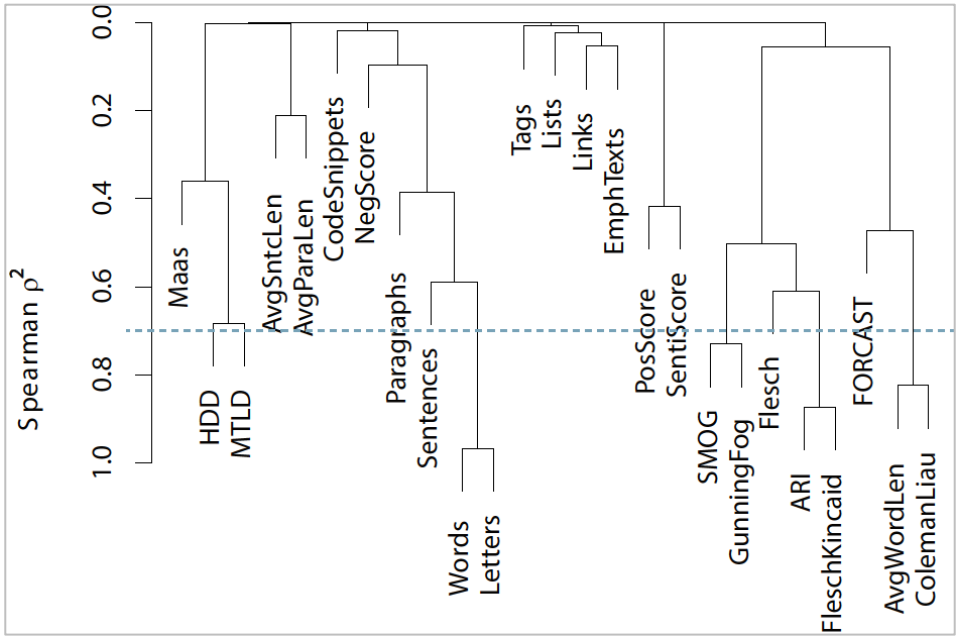


❖ Step 4: Rank Features by Importance

- Z-Score: mean decrease accuracy.

Results

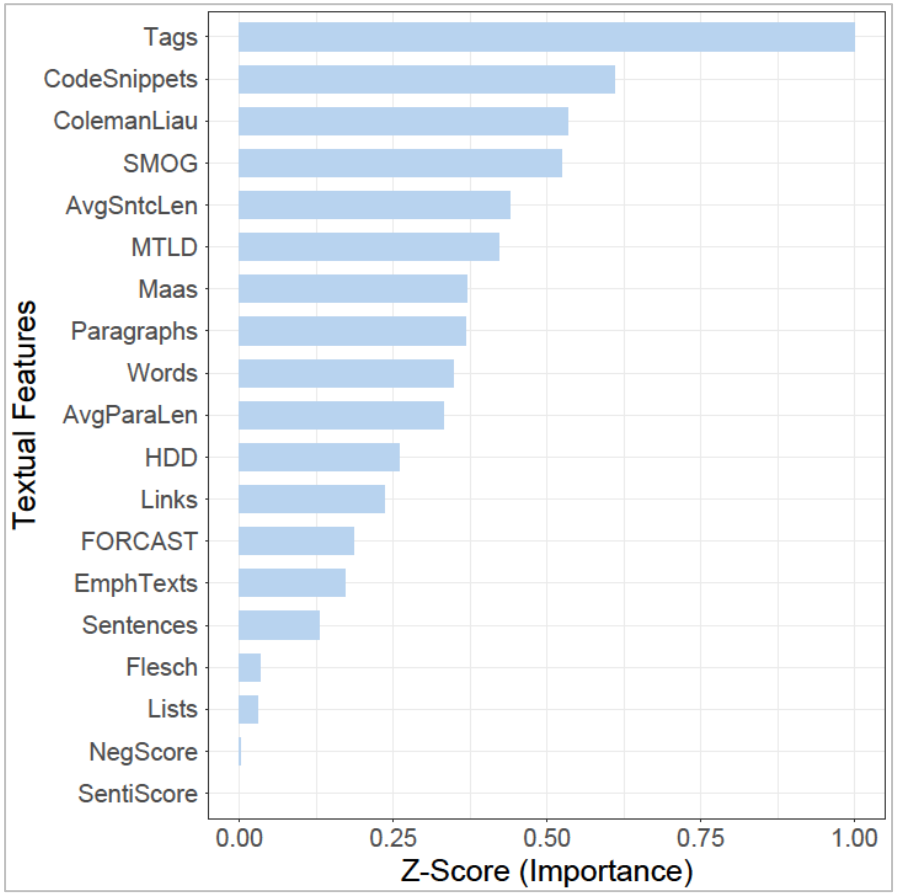
- Correlation Analysis and Redundancy Analysis
  - Remove Letters, GunningFog, ARI, and AvgWordLen.
  - Remove PosScore and FleschKincaid.



The Result of the Correlation Analysis

Empirical Findings

- Boruta Algorithm
  - 19 of the 25 selected features have relation with the question quality.
  - The number of tags is identified as the most discriminative one.
  - There is only a weak correlation between the question quality and the sentiment-related factors.



Rank Features by Importance

Empirical Findings

❖ Implications

- A checklist for SO members to optimize their questions.

Checklist for Making a High-Quality Question

Dimension	Feature	Rel.	Suggestion
Element	Tags	+	Include all relevant tags.
	CodeSnippets	+	Provide code snippets.
	Links	+	Present well-sourced facts.
	EmphTexts	+	Make the content skimmable
	Lists	+	with emphasized texts and lists.
Readability	ColemanLiau	+	Check the content readability.
	SMO	+	Use simple language (if possible,
	FORCAST	+	aim for a readability degree
	Flesch	+	below the 10th-grade level).

Empirical Findings

❖ Implications

- A checklist for SO members to optimize their questions.

Checklist for Making a High-Quality Question

Dimension	Feature	Rel.	Suggestion
Size	AvgSntcLen	-	Keep sentences short.
	Paragraphs	+	Break content into paragraphs.
	Words	+	Provide a relatively detailed
	AvgParaLen	+	description, yet keep the
	Sentences	+	content to the point.
Lexical Diversity	MTLD	+	Examine the lexical diversity.
	Maas	-	Aim for a low value as the metric
	HDD	+	is an indicative of text difficulty.
Sentiment	NegScore	-	Detect the sentiment polarity.
	SentiScore	-	Use neutral wording.

Threats to Validity

- Dataset Selection
- The indicator of the question quality
- The choice of textual features
- The use of SentiStrength

Conclusions

- An empirical study was conducted to provide insights into the textual features of high-quality questions.
- A set of practical suggestions was presented for guiding SO members on how to optimize their questions.