Argument Quality Prediction for Ranking Documents

Moritz Plenz, Raphael Buchmüller and Alexander Bondarenko
From Team Renji Abarai
Motivation

Task

Retrieval
Motivation

Task

Domain

"standard"

"argumentative"
Motivation

Task

Domain

Relevance

Retrieval

"standard"

"argumentative"

Documents

Documents
Motivation

Task
- Retrieval

Domain
- "standard"
- "argumentative"

Relevance
- Documents
- Documents

Quality
- Text
- Arguments
Motivation

Task

Domain

"standard"

"argumentative"

Relevance

Documents

Documents

Quality

Text

Arguments

Focus on Argument Quality
Initial Retrieval

Relevance: 3
Quality: 2
Stance: None

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con
Initial Retrieval

Relevance: 3
Quality: 2
Stance: None

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con

Keep top-10 documents
Initial Retrieval

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 1
  - Quality: 1
  - Stance: Con

Re-ranking by Quality

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 1
  - Quality: 1
  - Stance: Con
Initial Retrieval

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 1
  - Quality: 1
  - Stance: Con

Re-ranking by Quality

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 1
  - Quality: 1
  - Stance: Con

Re-ranking by Quality also improves Relevance
Initial Retrieval

Relevance: 3
Quality: 2
Stance: None

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Quality

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 3
Quality: 2
Stance: None

Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Stance & Quality

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con

Relevance: 3
Quality: 2
Stance: None
**Initial Retrieval**

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 1
  - Quality: 1
  - Stance: Con

**Re-ranking by Quality**

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 3
  - Quality: 2
  - Stance: None

- Relevance: 1
  - Quality: 1
  - Stance: Con

**Re-ranking by Stance & Quality**

- Relevance: 2
  - Quality: 3
  - Stance: Pro

- Relevance: 1
  - Quality: 1
  - Stance: Con

- Relevance: 3
  - Quality: 2
  - Stance: None

None means no arguments, i.e. poor Quality
Initial Retrieval

• BM25F-based search engine ChatNoir

Should teachers get tenure?
Initial Retrieval

- BM25F-based search engine ChatNoir

Should teachers get tenure?

Remove Punctuation & custom Stopwords

Lemmatize

should teacher tenure
Initial Retrieval

• BM25F-based search engine ChatNoir

Should teachers get tenure?

Remove **Punctuation** & custom **Stopwords**

Keep **Argumentative Stopwords**

Lemmatize

should teacher tenure
How to predict Argument Quality?

Manual Features
How to predict Argument Quality?

Manual Features

Neural Embeddings
How to predict Argument Quality?

- Manual Features
- Neural Embeddings
- In-context Learning
Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length
Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

- Punctuation
- Numerics
- External links
- Uppercase
- ...

Length

Occurances
Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length

- Punctuation
- Numerics
- External links
- Uppercase
- ...

Occurrences

- Academic
- Profanity
- Vocabulary richness
- ...

Word lists
Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

- Punctuation
- Numerics
- External links
- Uppercase
- ...

- Academic
- Profanity
- Vocabulary richness
- ...

- Number of arguments
- Subjectivity
- Sentiment
- Readability
- ...

Length
Occurrences
Word lists
Complex
Quality prediction: Manual Features

- **Length**
  - Document
  - Paragraphs
  - Sentences
  - Words

- **Occurrances**
  - Punctuation
  - Numerics
  - External links
  - Uppercase
  - ...

- **Word lists**
  - Academic
  - Profanity
  - Vocabulary richness
  - ...

- **Complex**
  - Number of arguments
  - Subjectivity
  - Sentiment
  - Readability
  - ...

- Total of 32 features
Quality prediction: Automatic Features
Quality prediction: Automatic Features

Document

NLTK

Sentences

Document Embedding
Quality prediction: Automatic Features

Document → Sentences → Sentence Embeddings → Document Embedding

NLTK → INSTRUCTOR
Quality prediction: Automatic Features

Document

Sentences

Sentence Embeddings

Document Embedding
Quality prediction: Classifier

Features
- Manual
- Automatic
Quality prediction: Classifier

- Features
  - Manual
  - Automatic

- Classifier
  - 6 shallow classifiers
  - Trained for manual and automatic separately
Quality prediction: Classifier

Features
- Manual
- Automatic

Classifier
- 6 shallow classifiers
- Trained for manual and automatic separately

Meta Learner
- Aggregates 12 classifier outputs
Quality prediction: Classifier

- Features
  - Manual
  - Automatic

- Classifier
  - 6 shallow classifiers
  - Trained for manual and automatic separately

- Meta Learner
  - Aggregates 12 classifier outputs

- Prediction
Quality prediction: Classifier

- **Features**
  - Manual
  - Automatic

- **Classifier**
  - 6 shallow classifiers
  - Trained for manual and automatic separately

- **Meta Learner**
  - Aggregates 12 classifier outputs

- **Prediction**

- Data from Touché 2021 Task 1
- 3 Quality Labels: low, medium and high
- Cross-topic split
Your task is to predict in a given text the rhetorical argument quality, i.e. "well-writteness": (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc.
You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.
Your task is to predict in a given text the rhetorical argument quality, i.e. "well-writtenness": (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc.

You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.

[ More examples... ]

[ More correct labels... ]

[ Document ]

[ Predicted label ]
Quality prediction: ChatGPT

Your task is to predict in a given text the rhetorical argument quality, i.e., "well-writtenness":
(1) whether the document contains arguments and whether the argument text has a good style of speech,
(2) whether the text has a proper sentence structure and is easy to follow,
(3) whether it includes profanity, has typos, etc.
You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.

[ More examples... ]

[ More correct labels... ]

[ Document ]

[ Predicted label ]
Your task is to predict in a given text the rhetorical argument quality, i.e. "well-written": (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc.

You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.
Your task is to predict in a given text the rhetorical argument quality, i.e. "well-writteness": (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc. You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.

<table>
<thead>
<tr>
<th>Predicted label</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
</tr>
<tr>
<td>[ More examples... ]</td>
</tr>
<tr>
<td>[ More correct labels... ]</td>
</tr>
<tr>
<td>[ Document ]</td>
</tr>
<tr>
<td>[ Predicted label ]</td>
</tr>
</tbody>
</table>
Stance prediction: ChatGPT

Instruction

Few-shot (4 examples)

Prediction

Document & Query
Initial Retrieval

Relevance: 3
Quality: 2
Stance: None

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Quality

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 3
Quality: 2
Stance: None

Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Stance & Quality

Relevance: 2
Quality: 3
Stance: Pro

Relevance: 1
Quality: 1
Stance: Con

Relevance: 3
Quality: 2
Stance: None
Results – Stance

![Stance Graph](image-url)

- ChatGPT: Macro F1
- Flan-T5: Macro F1
Results – Stance

![Stance Chart]

- **ChatGPT**
  - Macro F1: 0.6
  - Stance: 0.6

- **Flan-T5**
  - Macro F1: 0.2
  - Stance: 0.2

![Stance (Binary) Chart]

- **ChatGPT**
  - Macro F1: 0.7

- **Flan-T5**
  - Macro F1: 0.4
Results – Quality Classification

[Bar chart showing macro F1 scores for Feature-based, Embedding, Meta Learner, and ChatGPT.]

- Feature-based: Around 0.30
- Embedding: Around 0.30
- Meta Learner: Around 0.40
- ChatGPT: Above 0.40
Results – Quality Ranking
Results – Quality Ranking

![Diagram showing nDCG@10 for different methods: Baseline, Feature-based, ChatGPT, and Baseline again. The x-axis represents the methods, and the y-axis represents nDCG@10. The methods are compared with Renji Abarai and Puss in Boots.]
Results – Quality Ranking

Task baseline outperforms all submissions
Results – Quality Ranking

![Graph showing nDCG@10 for different models: Baseline, Feature-based, ChatGPT, Stance, Baseline, ChatGPT, Stance. The graph compares two models: Renji Abarai and Puss in Boots.]
Results – Quality Ranking

Re-ranking improves task baseline
Results – Relevance Ranking

![Graph showing nDCG@10 for different methods: Baseline, Feature-based, ChatGPT, Slance, Baseline, ChatGPT, and Stance. The x-axis represents the methods, and the y-axis represents the nDCG@10 scores. The graph compares two entities: Renji Abarai and Puss in Boots.]
Results – Relevance Ranking

Most improvement from **Stance**
Conclusion

• **BM25F** is a strong baseline
• **Quality** consistently improves with re-ranking
• **Relevance** improves mostly from considering the stance in re-ranking
# Appendix – Table

<table>
<thead>
<tr>
<th>Configuration (run)</th>
<th>nDCG@10</th>
<th></th>
<th>macro F1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quality</td>
<td>Relevance</td>
<td>Stance</td>
<td>Stance bin.</td>
</tr>
<tr>
<td>BL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatNoir [9] / Flan-T5 (stance) [38]</td>
<td>0.831</td>
<td>0.834</td>
<td>0.203</td>
<td>0.432</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stance_ChatGPT</td>
<td>0.815†</td>
<td>0.744</td>
<td>0.599</td>
<td>0.780</td>
</tr>
<tr>
<td>stance-certainNO_ChatGPT</td>
<td>0.811†</td>
<td>0.746</td>
<td>0.604(*)</td>
<td>0.783(*)</td>
</tr>
<tr>
<td>ChatGPT_mmGhl</td>
<td>0.789</td>
<td>0.718</td>
<td>0.599</td>
<td>0.780</td>
</tr>
<tr>
<td>ChatGPT_mmEQhl</td>
<td>0.789</td>
<td>0.718</td>
<td>0.599</td>
<td>0.780</td>
</tr>
<tr>
<td>meta_qual_prob</td>
<td>0.774</td>
<td>0.697</td>
<td>0.599</td>
<td>0.780</td>
</tr>
<tr>
<td>meta_qual_score</td>
<td>0.771</td>
<td>0.712</td>
<td>0.599</td>
<td>0.780</td>
</tr>
<tr>
<td>baseline</td>
<td>0.766</td>
<td>0.708</td>
<td>0.599</td>
<td>0.780</td>
</tr>
</tbody>
</table>
Appendix – Quality Classification Detailed
Appendix – Quality Classification ChatGPT

![Classification Matrix]

- **True Class**
- **Predicted Class**

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low</strong></td>
<td>0.54</td>
<td>0.16</td>
<td>0.29</td>
</tr>
<tr>
<td><strong>Medium</strong></td>
<td>0.36</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>High</strong></td>
<td>0.15</td>
<td>0.3</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Appendix – Quality Classification with Stance ChatGPT

![Heatmap showing classification results.](image)
Appendix – Quality Ranking Detailed
Appendix – Relevance Ranking Detailed