Reddit comments to post Relevance Analysis

Rajesh Kumar Reddy Avula
Indiana University Bloomington
rajavula@iu.edu

Sai Sindhura Kollepara
Indiana University Bloomington
skollep@iu.edu

Shoukath Ali Shaik
Indiana University Bloomington
shshaik@iu.edu

Vineetha Maddikunta
Indiana University Bloomington
vimadd@iu.edu

May 1, 2024

GitHub Repository: https://github.iu.edu/Luddy-B565-SP24/rajavula-skollep-shshaik-vimadd

Abstract

This paper addresses to analyze the relevance of the users' comments to the Reddit posts, and aims to give a rating on a scale of 1 to 5, 5 being the highest relevance and 1 being the lowest comment. Our approach includes multiple steps, beginning with the collection of data through a Reddit developer account and pre-processing to get rid of deleted/removed comments, spams, HTML tags, URLs, punctuations and emojis. We then apply TF-IDF vectorization for the post titles and comments and create embeddings using various models including Latent Dirichlet Allocation (LDA), Long Short-Term Memory (LSTM) neural networks, pretrained sentence transformers like GPT, BERT and MiniLM, and the hybrid models with the combinations of all individual models then clustering these embeddings using KMeans. Rating is assigned to the clusters formed on a scale of 1 to 5 using a custom relevant metric which considers weighted User Engagement Factors like Score, and Reply count, Cluster Coherence, Semantic Similarity and Comment Polarity. Our findings display the productiveness of this methodology in accurately identifying relevant comments to the Reddit posts.

Keywords : Relevance, vectorization, embeddings, clustering, topic modelling, pre-trained sentence transformers, hybrid models.

1 Introduction

Online social media forums such as Quora, Reddit, Stack Exchange etc., are popular websites for discussions, where people share their opinions and exchange information on various topics. Finding only the relevant comments which provide meaningful insights to the question being asked within the hundreds of comments being shared becomes a challenge.

The main focus area of this study is to come up with an efficient method to identify and classify these comments based on their relevance with respect to the post. By achieving this, we can enhance user experience by promoting meaningful discussions, aiding moderators in identifying relevant comments efficiently.

The motivation behind this project comes from the ever-growing demand of online interaction in framing people's viewpoints, influencing their decision-making approach and encouraging community participation. As now-a-days online platforms continue to grow into one of the primary sources for communication and information, finding effective approaches to filter, analyze and organize the content created by users becomes way more important.

This paper specifically focuses on Reddit, a well-liked and in-demand social media platform recognized for its wide spectrum of communities and discussions.

By focusing on Reddit, we aim to tackle this challenge by proposing a solution for identifying relevant comments to the Reddit posts, comments being giving rating on a scale of 1 to 5, 5 being highest and 1 being the lowest which can be applied to other similar online platforms.

To summarize, the main goals of this project are:

1. To find an effective approach for classifying users' comments into categories with respect to the question being asked in Reddit.

2. To implement the recommended approach for identifying and classifying the comments into five categories, taking their relevance rating based on our custom relevance metric into account from highest to lowest.

This study paves a path for future advancements for enhancing user experience by promoting meaningful discussions such as considering multi-modal relevance assessments etc.

2 Related Work

1. Integration of Topic Modeling with Neural Networks
   Traditional topic modeling techniques, particularly Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan, have been pivotal in identifying latent topic distributions within large text datasets. While LDA has been effective for topic discovery, its reliance on the bag-of-words assumption limits its ability to grasp deeper semantic meanings. This limitation has ignited interest in hybrid models that combines the thematic strength of LDA with the deeper understanding offered by deep learning techniques [Li et al., 202][1].

2. Enhancements through Deep Learning
   Integrating LDA outputs with neural networks has shown a significant potential and this hybrid approach uses topic distributions as features within neural networks to improve tasks like text classification, sentiment analysis, and content recommendation, resulting in more robust text data analysis [Shahbazi & Byun, 2021][2]. Additionally, MiniLM—a streamlined version of large language models—offers a compact yet powerful framework for integrating deep semantic representations into these hybrid systems, enabling faster processing.

3. Application in Question Answering and Document Retrieval
   LDA combined with neural networks or pre-trained models is useful in question answering and document retrieval tasks. By incorporating topic proba-

bilities from LDA into deep learning models, it helps in improving the accuracy of responses[Coughlin et al., 2017][3].

3 Methodology

3.1 Environment Details

The environment details utilized while performing this project are as follows:

- Python Version: 3.10.12
- Operating System: Linux #1 SMP PREEMPT_DYNAMIC
- Machine: x86_64
- CPU Cores: 1 Physical, 2 Logical
- Total Memory: 12.67 GB
- No GPU available.
- name: "/device:CPU:0"
- device_type: "CPU"
- Total Disk Space: 107.72 GB

3.2 Dataset

Beginning with the dataset acquisition for this problem statement, we gathered posts-comments utilizing Reddit's developer access. Facilitated with the client ID and secret credentials, we accessed the post's URLs. Our dataset comprises 10 distinct questions, accompanied by their corresponding comments, scores, and replies count where 'score' denotes the ratio of upvotes to downvotes as provided by Reddit itself. However, we encountered a challenge during this phase, wherein the API call yielded a 429 error, restricting access to questions with substantial comment volumes. Consequently, we proceeded with posts featuring a more moderate level of engagement. Each of these questions have been organized into individual Excel sheets and were consolidated within a folder for easy access.

3.3 Pre-processing

Upon assembling the dataset, our next step is to take pre-processing measures to ensure data quality and uniformity. We began by addressing any null comments encountered, and handling comments that were either deleted by users or removed. Since Reddit supports diverse content-sharing capabilities, we encountered numerous URLs redirecting to multimedia content such as gifs or YouTube videos. To maintain focus on textual analysis, we removed such URLs from the dataset. Furthermore, we handled emojis, special characters and punctuation as well.

We employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique for vectorizing both comments and questions which gives us robust numerical representations of the text data. TF-IDF, renowned for its usage in information retrieval tasks, emphasizes the terms that are quite unique or rare to specific documents while de-emphasizing commonly used terms prevalent across the corpus. This ensures that the resultant vector representations capture the essence of each comment and question.

3.4 Building a model

The design approach for this has three key stages:
The initial step involves generating embeddings of the text vectors. We explored

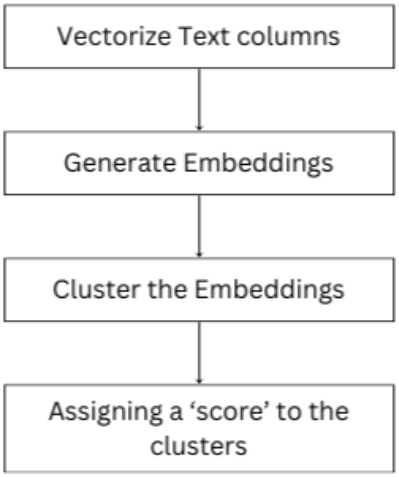

Figure 1: Design Flow.

various techniques, including Latent Dirichlet Allocation (LDA) for topic modeling, as well as leveraging pre-trained sentence transformers such as GPT, BERT, BERT Quora, and MiniLM. Additionally, neural network architectures like Long Short-Term Memory (LSTM) networks were utilized as well. We also delved into hybrid models, combining LDA as the primary model with others serving as secondary models. While considering the hybrid models, we have combined the first model embeddings with the second model embeddings to form a new numeric representation of the text vectors.

Subsequently, we utilized a KMeans clustering model to partition the generated embeddings into five clusters. This choice of five clusters corresponds to the envisioned relevance rating scale, ranging from 1 to 5.

Following clustering, our focus shifted to assigning a value to each cluster to elucidate its relevance. To accomplish this, we devised a metric of our own. This metric factors in several aspects, including the semantic similarity between comments and questions using cosine similarity, cluster coherence via pairwise similarity, sentiment analysis of individual comments, comment score, and reply count. Utilizing this metric, we determined a relevance rating for each cluster, thereby establishing their hierarchical order.
The formula used in this project for calculating score is as follows:

$$relevance\_score = mean\_score + (0.25 \times mean\_reply\_count)$$

$$+mean\_sentiment + np.mean(semantic\_similarity) + coherence$$

4 Results

LDA exhibited commendable performance, and so did BERT Quora and MiniLM. Notably, the combination of LDA with other models showcased even better performance.

Considering the question : 'What do you all think is the best IPhone 15 pro max color?'
The results provided by hybrid models:

1. LDA and Bert :

	4	3	2	0	1
0	in order for me is naturalwhiteblackblue the n...	blue i have blue bought my wife the natural on...	black definitely	1 and 3	to me the black phones make the bezels look la...
1	i ordered black titanium not received it yet b...	this is true i almost got one in natural for t...	honestly i wouldnt fault anyone for choosing a...	batman	so let me understand the paint or whatever on ...
2	definitely not the natural titanium everyone s...	i got the blue and i love the look but i did a...	i got black because thats the only color they ...	natural or black thumbsup	you can barely see fingerprints on it and any ...
3	theres something about the white and natural t...	de gustibus non disputandum est whatever you ...	i got the white one i tend to go to with white...	ditz is away	press on the back glass with your thumbs then ...

```
relevance_scores_h1 = calculate_relevance_scores(question1df,question_combined1,cluster_embeddings_h1,topics_hybrid1,'lda')
print("Relevance Scores for Each Cluster:", relevance_scores_h1)
```

```
Relevance Scores for Each Cluster: {4: 1.0, 3: 0.9406273142563776, 2: 0.6422345189690409, 0: 0.0, 1: 0.48989442687790136}
```

Figure 2: LDA + Bert model clusters and scores.

2. LDA and BertQuora :

	3	2	0	4	1
0	in order for me is naturalwhiteblackblue the n...	blue i have blue bought my wife the natural on...	black definitely	1 and 3	to me the black phones make the bezels look la...
1	i ordered black titanium not received it yet b...	this is true i almost got one in natural for t...	honestly i wouldnt fault anyone for choosing a...	batman	so let me understand the paint or whatever on ...
2	definitely not the natural titanium everyone s...	i got the white one i tend to go to with white...	i got black because thats the only color they ...	ditz is away	you can barely see fingerprints on it and any ...
3	theres something about the white and natural t...	i got the blue and i love the look but i did a...	1 natural 2 black 3 blue 4 white	thank you for your donation	press on the back glass with your thumbs then ...

```
relevance_scores_h2 = calculate_relevance_scores(question1df,question_combined2,cluster_embeddings_h2,topics_hybrid2,'lda')
print("Relevance Scores for Each Cluster:", relevance_scores_h2)
```

```
Relevance Scores for Each Cluster: {3: 1.0, 2: 0.7884593376142386, 0: 0.6944928810939822, 4: 0.0, 1: 0.9203734957925688}
```

Figure 3: LDA + BertQuora model clusters and scores.

3. LDA and MiniLM :

	2	1	3	0	4
0	in order for me is naturalwhiteblackblue the n...	blue i have blue bought my wife the natural on...	black definitely	1 and 3	to me the black phones make the bezels look la...
1	i ordered black titanium not received it yet b...	this is true i almost got one in natural for t...	honestly i wouldnt fault anyone for choosing a...	that blue smilingfacewithhearteyes	so let me understand the paint or whatever on ...
2	definitely not the natural titanium everyone s...	i got the white one i tend to go to with white...	i got black because thats the only color they ...	batman	you can barely see fingerprints on it and any ...
3	theres something about the white and natural t...	i got the blue and i love the look but i did a...	1 natural 2 black 3 blue 4 white	ditz is away	press on the back glass with your thumbs then ...

```
relevance_scores_h3 = calculate_relevance_scores(question1df,question_combined3,cluster_embeddings_h3,topics_hybrid3,'lda')
print("Relevance Scores for Each Cluster:", relevance_scores_h3)
```

```
Relevance Scores for Each Cluster: {2: 1.0, 1: 0.6485419306679798, 3: 0.3249707490111305, 0: 0.0, 4: 0.549980451262985}
```

Figure 4: LDA + MiniLM model clusters and scores.

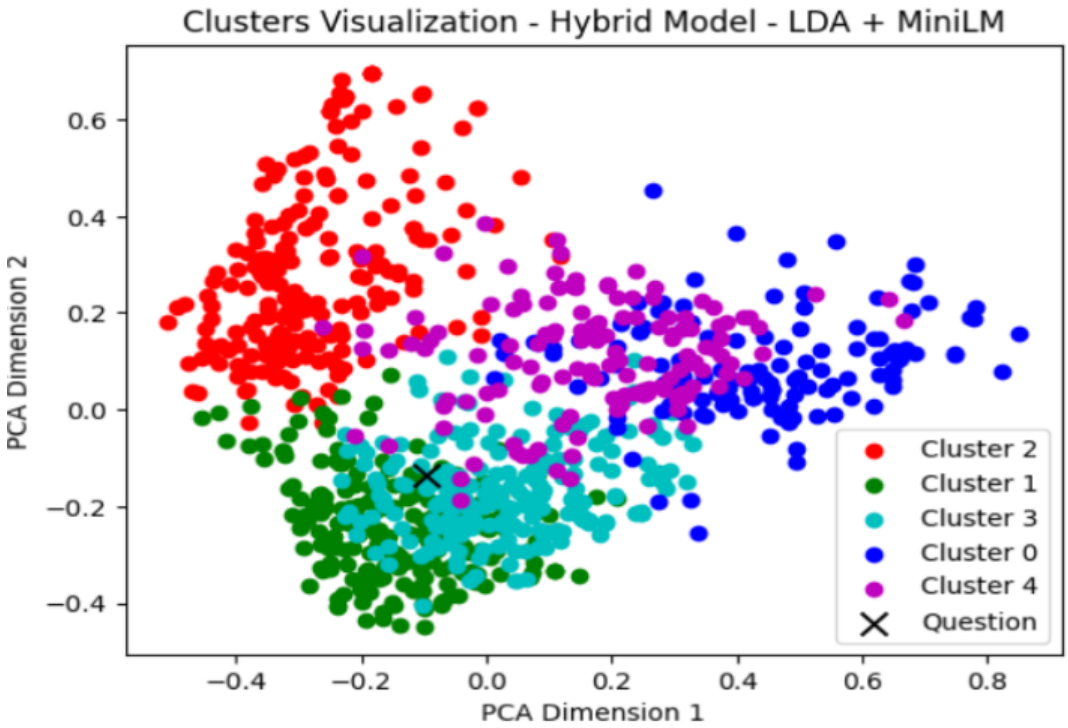

Figure 5: LDA + MiniLM model clusters using T-SNE.

If we simply turn to clusters to calculate the scores based on any distance metrics, it would not be enough to properly represent the relevance. And this is where we also consider the scores generated by our metric to evaluate the accuracy of the model.

Given the differences in embeddings derived from each model, it follows that the resultant clusters exhibit variations as well. Upon visual examination, it becomes evident that among the hybrid models tested, MiniLM yields better clusters.

5 Conclusion

Moving forward, the refinement of the relevance metric remains a viable platform for enhancement. This could involve introducing weighted considerations for the factors utilized and incorporating additional factors as necessitated by the task at hand. It's important to note that, presently, our project lacks a formal evaluation technique for assessing model performance. Instead, model effectiveness has been discerned through visual inspection of results. Future endeavors may benefit from the implementation of evaluation methodologies to validate model accuracy.

An additional aspect worth considering is the multi-modal content on Reddit. We have excluded such data in this project during pre-processing stages but, incorporating such content into this project by extracting and analyzing text data from these posts and comments could significantly enhance its efficiency. By integrating multi-modal data, we broaden the scope of analysis, potentially uncovering valuable insights and enriching the overall findings of the project.

References

[1] Li, X., Pang, J., Mo, B., & Rao, Y., *Hybrid neural networks for social emotion detection over short text*, *Social Network Analysis and Mining*, vol. 12, no. 1, p. 139, doi: 10.1007/s13278-022-00957-x,2022.

[2] Shahbazi, Z., & Byun, Y., *Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning.*, *Journal of Intelligent & Fuzzy Systems* doi: 10.3233/JIFS-191690, IOS Press, 2020.

[3] Coughlin, R., Coetsier, J.-C., & Jiamthapthaksin, R., *Integrating Labeled Latent Dirichlet Allocation into sentiment analysis of movie and general domains*, *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pp. 18–22, 2017.

[4] Murakami, R., & Chakraborty, B., *Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts*, *Sensors*, vol. 22, no. 3, pp. 852, doi: 10.3390/s22030852, January 2022.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 [cs.CL], 24 May 2019.

[6] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M., *MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*, arXiv:2002.10957 [cs.CL], 6 Apr 2020.

[7] Jelodar, H., Wang, Y., Yuan, C., & Feng, X., *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*, November 2017.

[8] Yang, D., Zhu, D., Gai, H., & Wan, F., *Semantic Similarity Calculating based on BERT*, *Journal of Electrical Systems*, vol. 20, no. 2, pp. 73–79, doi: 10.52783/jes.1099, License: CC BY-ND 4.0, April 2024.

[9] Khodke, S., & Dhande, S., *A Systematic Review of Similar Questions Retrieval Approaches*, *Asian Journal of Convergence in Technology*, vol. 9, no. 2, pp. 29–42, doi: 10.33130/AJCT.2023v09i02.006, August 2023.

[10] Olaoye, F., & Potter, K., *Transfer Learning and its Role in Machine Learning*, *Technology*, Authors: Favour Olaoye, Kaledio Potter, March 2024.

[11] Tan, P.-N., Steinbach, M., & Kumar, V., *Introduction to Data Mining (2nd Edition)*, 2nd Edition, Pearson, Publication date: January 4, 2018.