Beyond Relevance: IntenRank for Pair-Wise Semantic Intensity Ranking in Information Retrieval and Evaluation

Anonymous ACL submission

Abstract

Information retrieval systems primarily focuses on semantic relevance to rank text on how closely a document aligns with a query. This approach has being adopted in many areas including e-commerce, survey analysis, and Retrieval-Augmented Generation (RAG) for large language models (LLMs). While it is effective in many cases, it fails to capture semantic intensity—the degree of emotion, emphasis, or strength stressed in text. Semantic intensity provides insights across many dimensions, assisting applications like survey analysis, product reviews, and information retrieval systems. Our experiment results demonstrate that having semantic relevance alone is insufficient for tasks requiring effective intensity-based rankings. We introduce IntenRank, a novel pair-wise ranking model that evaluates the relative semantic intensity of texts with respect to a given query. Evaluated on e-commerce review datasets, IntenRank demonstrates significant improvements over traditional relevance-based ranking, achieving notable gains in ranking correctness, especially in scenarios where capturing differences in semantic intensity is critical.

1 Introduction

Traditional information retrieval systems have primarily focused on semantic relevance when ranking documents and responses.This approach is effective in numerous applications, including e-commerce, survey analysis, and Retrieval-Augmented Generation (RAG) for large language models (Lewis et al., 2020; Karpukhin et al., 2020). Recent advancements in RAG have further improved performance through reranking techniques (Wang et al., 2023a; Tay et al., 2022).

However, these systems fall short when the strength or intensity of meaning becomes crucial. For example, in tasks such as product review ranking or customer feedback prioritization, understanding not just semantic meaning but how strongly certain aspects, like product adaptability and product durability, are expressed is essential. Current systems often overlook this critical dimension (de Marneffe et al., 2013), which plays a key role in areas such as customer feedback and sentiment analysis (Zhang et al., 2021a).

The concept of semantic intensity, first explored in visual relevance (Liu et al., 2006), is crucial in textual information retrieval for applications requiring understanding of message strength, especially in subjective domains like customer reviews. While semantic relevance helps retrieve similar texts, it's insufficient for tasks needing deeper comprehension of semantic intensity (de Marneffe et al., 2013). Applications like customer feedback prioritization and sentiment analysis require capturing nuanced intensity signals, which traditional models often overlook (Zhang et al., 2021b; Ruppenhofer and Brandes, 2015).

This paper introduces IntenRank, a novel pair-wise ranking model for semantic intensity in text. It addresses limitations in current retrieval systems by capturing and ranking intensity levels across multiple dimensions. IntenRank can enhance RAG systems by leveraging advanced LLMs like LLaMA 3 (Dubey et al., 2024), improving accuracy and contextual appropriateness in responses. This approach tackles the challenge of capturing expression strength and intensity in current search and retrieval paradigms.

2 Related Work

Semantic Intensities and Information Retrieval. Semantic intensity in text has been explored in various contexts, from visual relevance (Liu et al., 2006) to language intensity scales (de Marneffe et al., 2013; Ruppenhofer and Brandes, 2015). Traditional IR models like BM25 (Beaulieu et al., 1997) indirectly captured emphasis through term frequency and document length. However, these

approaches lack the nuanced and broad understanding of semantic strength/intensity that modern techniques aim to achieve.

Recent Advances in IR and E-commerce. Information retrieval systems have evolved significantly, especially in e-commerce. Recent advancements include deep learning-based pairwise ranking (Zhang et al., 2021b), sentiment-aware product search (Zhang et al., 2021a), semantic item recommendation (Hou et al., 2024) and semantic deep neural networks for sentiment analysis (Liu et al., 2022). Neural models like SPLADE (Formal et al., 2021) and ColBERT (Khattab and Zaharia, 2020) have further improved retrieval performance. Notably, e-commerce platforms are increasingly adopting these advanced models, including ColBERT and similar machine learning approaches, for their information retrieval systems.

Advanced Ranking Models. Novel ranking paradigms have emerged to address complex retrieval challenges. Recent advancements include MatRank for text re-ranking using latent preference matrices (Zhang et al., 2022), instruction distillation for zero-shot ranking with large language models (Wang et al., 2023b), and modular re-rankers for long document ranking (Wu et al., 2023). Pairwise ranking models have gained popularity, with research focusing on approximate techniques for scalability (Heckel et al., 2018) and efficient methods for pairwise comparisons (Wauthier et al., 2013).

Language Models in IR. The introduction of BERT (Devlin et al., 2018) marked a pivotal moment in Natural Language Processing (NLP) and Information Retrieval (IR), revolutionizing various tasks. Building on this foundation, Lin et al. (Lin et al., 2021) provided a comprehensive overview of transformer applications in text ranking. Kaplan et al. (Kaplan et al., 2020) explored scaling laws for neural language models, suggesting potential improvements through increased model size and data volume.

These advancements have enhanced the capture of relevance and semantic relationships. However, they often overlook the nuanced aspects of semantic intensity. *IntenRank* addresses this gap by focusing on the strength of sentiment in ranking tasks, offering a more comprehensive approach where the degree of expression is as crucial as the content itself.

3 Methods

3.1 IntenRank Overview

IntenRank is a novel pairwise ranking model that captures semantic intensity in text. Unlike traditional relevance-based models, it ranks texts based on the strength of emotional or emphatic signals relative to a query. This approach improves ranking for tasks where intensity is crucial, such as prioritizing customer feedback or analyzing product reviews.

3.2 Intensity Features.

To capture the intensity of textual content, in this case the product reviews, we introduce five key dimensions: *Repurchasement*, *Durability*, *Satisfaction*, *Comparison*, and *Adaptability*. These dimensions are embedded as contextual signals during fine-tuning to ensure that the model learns to rank texts not just based on relevance but also on the strength of these features. Each feature is labeled with a value ranging from 0 to 9 using GPT-4o, a variant of the GPT-4 large language model (OpenAI, 2023).

3.3 Model Architecture and Training

The base model of *IntenRank* is built on DistilBERT (Sanh et al., 2019), a lighter and faster version of BERT. We fine-tuned DistilBERT using the Hugging Face Transformers library (Wolf et al., 2020) to predict pairwise rankings based on semantic intensity. During training and inference, we structure the input to DistilBERT as a concatenation of the query, review1, and review2, separated by [SEP] tokens. This format can be represented as: "query [SEP] review1 [SEP] review2". This input structure allows the model to learn the relationships between the query (which represents the intensity feature) and the pair of reviews being compared, enabling effective pairwise ranking based on semantic intensity.

3.4 Pairwise Ranking Objective

The model is trained as a pairwise classifier, where the goal is to determine which text in a given pair has higher intensity. For each pair of texts, the model outputs logits corresponding to two possible labels: 0 (indicating that the second text has higher intensity) and 1 (indicating that the first text has higher intensity).

The model applies *argmax* on the logits to predict which text has higher intensity. The loss func-

Feature	Model	Avg Precision@5	Avg Precision@10	Avg Kendall's Tau
Repurchase	Baseline	0.20	0.13	0.0724
Repurchase	IntenRank	0.73	0.67	0.7680
Durability	Baseline	0.00	0.23	0.0994
Durability	IntenRank	0.40	0.57	0.6615
Satisfaction	Baseline	0.13	0.13	0.0049
Satisfaction	IntenRank	0.73	0.73	0.7764
Comparison	Baseline	0.20	0.30	0.1625
Comparison	IntenRank	0.67	0.60	0.7179
Adaptability	Baseline	0.00	0.23	0.0020
Adaptability	IntenRank	0.27	0.40	0.3776

Table 1: Average Performance of Baseline and IntenRank Models across 5 Features

tion used is a standard CrossEntropyLoss, which encourages the model to correctly classify the relative intensity between the two texts. The CrossEntropyLoss is defined as:

$$L(\theta) = -\sum_{(i,j)\in P} [y_{ij} \log(\hat{y}_{ij})]$$

where P is the set of all pairs (i, j), with i representing the first text and j representing the second text. y_{ij} is the true label for the pair, where 1 indicates the first text has higher intensity and 0 indicates the second text has higher intensity. $\hat{y}_{ij}$ is the predicted probability that the first text has higher intensity than the second.

3.5 Baseline Model

To evaluate the effectiveness of IntenRank, we implement a baseline model that focuses on semantic relevance while attempting to capture specific semantic intensity features. This baseline utilizes DistilBERT (Sanh et al., 2019) for semantic similarity ranking, employing cosine similarity between text embeddings as the ranking criterion. We incorporate the HyDE (Hypothetical Document Embeddings) technique (Gao et al., 2022) to generate query embeddings that represent specific semantic intensity features, allowing us to attempt ranking reviews by particular intensity aspects using vector similarity.

3.6 Training and Hyperparameters.

IntenRank is trained using a batch size of 16, with a learning rate of 5e-5, over 1 epoch. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with weight decay to fine-tune the model on the dataset. Training is performed on an AWS g5.xlarge instance equipped with NVIDIA A10 GPU, and the model is optimized to minimize the pairwise ranking loss.

4 Experiments

To evaluate the performance of IntenRank, we conducted a series of experiments comparing it against a baseline model and using established metrics for ranking tasks.

4.1 Experimental Setup

4.1.1 Dataset

For our experiments, we used the Amazon All-Beauty product reviews dataset (Hou et al., 2024). We selected 20,000 product reviews from this dataset. To create our training set, we generated 100,000 paired reviews from these 20,000 reviews, excluding 10,000 reviews that were set aside for testing. This approach allowed us to create a robust training set for our pairwise ranking model while maintaining a separate test set for evaluation.

The ranking of paired reviews was computed by comparing one of the intensity feature scores labeled using GPT-4o (OpenAI, 2023).

4.1.2 Baseline Model

For comparison, we use a baseline model based solely on semantic relevance, which ranks texts without considering intensity. This baseline is implemented using DistilBERT (Sanh et al., 2019) for semantic similarity ranking. The baseline compares texts based on their embedding similarity, using cosine similarity as the distance metric:

4.1.3 Evaluation Metrics

To evaluate the performance of *IntenRank*, we use the following metrics:

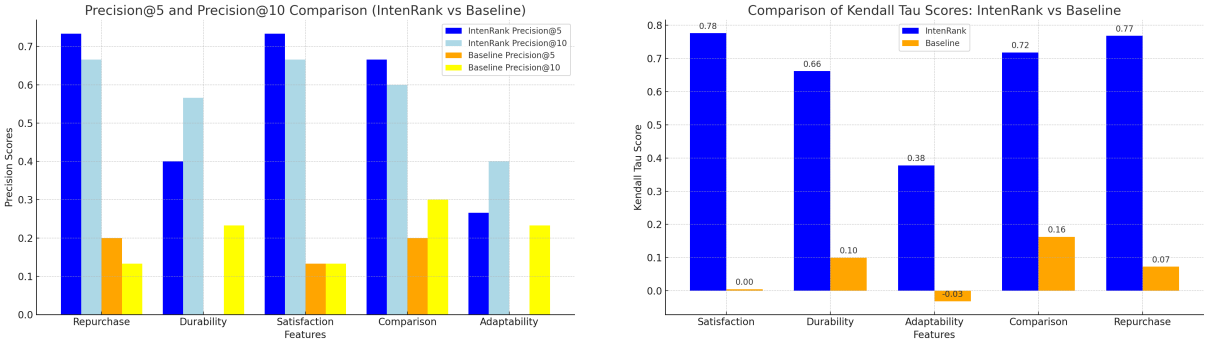

Figure 1: Comparison of IntenRank and Baseline on Precision and Kendall Tau Metrics. Left: Precision Scores (P@5 and P@10) for IntenRank vs Baseline. Right: Kendall Tau Scores for IntenRank vs Baseline.

- **Precision at K (P@K)**: This metric measures the proportion of relevant items among the top K items in the ranked list. We use IntenRank model compute ranking by pairwise comparing all reviews. We specifically use Precision at 5 (P@5) and Precision at 10 (P@10) to evaluate the model's performance at different depths of the ranking.

- **Kendall's Tau** (τ): This rank correlation coefficient measures the ordinal association between the predicted rankings and ground truth rankings. It is defined as:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

where C is the number of concordant pairs, D is the number of discordant pairs in the rankings, T_x is the number of ties in the first ranking, and T_y is the number of ties in the second ranking.

4.2 Results

The performance of IntenRank compared to the baseline model is summarized in Table 1 and visualized in Figure 1. Figure 1 shows the Precision at K scores (P@5 and P@10) and Kendall's Tau scores across all five features.

IntenRank consistently outperforms the baseline in precision across all features. It achieves notably high P@5 and P@10 scores for *Repurchase* (0.73 and 0.67) and *Satisfaction* (0.73 for both), compared to the baseline's lower scores (0.20 and 0.13 for Repurchase, 0.13 for both in Satisfaction). For *Adaptability* and *Durability*, IntenRank shows more modest but still superior performance.

The Kendall's Tau scores further demonstrate IntenRank's superiority. It shows higher correlation with the true ranking across all features, with the largest gains in *Satisfaction* (τ = 0.7764), *Repurchase* (τ = 0.7680), and *Comparison* (τ = 0.7179). The baseline, in contrast, shows much lower scores, particularly for *Satisfaction* (τ = 0.0049) and *Adaptability* (τ = 0.0020).

5 Limitations

Here are some limitations of IntenRank:

1. Domain Specificity: Primarily trained on e-commerce reviews, potentially limiting generalization to other domains.

2. Feature Scope: Focuses on five specific intensity features, which may not cover all aspects of semantic intensity across different contexts.

3. Scalability: The pairwise ranking approach may face challenges with very large datasets or real-time scenarios.

We anticipate improved performance and broader applicability as more diverse training data becomes available in future work.

6 Conclusion

This paper introduces IntenRank, a novel pairwise ranking model designed to capture and rank semantic intensity in text. Our experiments demonstrate that IntenRank significantly outperforms traditional relevance-based ranking models across multiple intensity features in e-commerce review scenarios. By focusing on the strength of sentiment rather than just relevance, IntenRank addresses a critical gap in current information retrieval systems. Future work will focus on expanding IntenRank's applicability to diverse domains and integrating it with large language models.

References

Micheline Beaulieu, M Gatford, X Huang, Stephen E Robertson, S Walker, and P Williams. 1997. Okapi at trec-5. In *The Fifth Text REtrieval Conference (TREC-5)*, pages 143–165. NIST Special Publication.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:297–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. *arXiv preprint arXiv:2107.05720*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. 2018. Approximate pairwise ranking. In *International Conference on Machine Learning*, pages 1957–1966. PMLR.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of NeurIPS*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*.

Ying Liu, Dengsheng Zhang, and Guojun Lu. 2006. Semantic intensity: a measure of visual relevance. *Information Sciences*, 176(19):2867–2881.

Yiyi Liu, Chengxiang Jiang, and Huimin Zhao. 2022. Sentiment analysis of consumer reviews using deep learning. *Sustainability*, 14(17):10844.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774.

Josef Ruppenhofer and Jasper Brandes. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2022. Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, 56(1):1–27.

Yujia Wang, Zhengliang Jiang, Yueyang Xu, Yixuan Xu, Weizhu Yin, Dawei Yin, Pengjie Ren, and Maarten de Rijke. 2023a. Don't forget to connect! improving rag with graph-based reranking. *arXiv preprint arXiv:2303.05968*.

Yujia Wang, Zhengliang Jiang, Yueyang Xu, Yixuan Xu, Weizhu Yin, Dawei Yin, Pengjie Ren, and Maarten de Rijke. 2023b. Instruction distillation makes large language models efficient zero-shot rankers. *arXiv preprint arXiv:2303.00453*.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. 2013. Efficient ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 109–117. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yixing Wu, Luyu Gao, Yixing Ren, Dawei Yin, and Jamie Callan. 2023. Long document re-ranking with modular re-ranker. *arXiv preprint arXiv:2303.05445*.

Yujia Zhang, Jing Xu, Changliang Jiang, Zhaochun Ren, Pengjie Dong, Zhumin Chen, and Maarten de Rijke. 2022. Matrank: Text re-ranking by latent preference matrix. *arXiv preprint arXiv:2203.15748*.

Yujie Zhang, Jing Xu, Changliang Jiang, Zhaochun Ren, Pengjie Dong, Zhumin Chen, and Maarten de Rijke. 2021a. Deep learning based sentiment aware ranking for e-commerce product search. In *International Conference on Web Information Systems Engineering*, pages 86–101. Springer.

Yujie Zhang, Jing Xu, Changliang Jiang, Zhaochun Ren, Pengjie Dong, Zhumin Chen, and Maarten de Rijke. 2021b. From semantic retrieval to pairwise ranking: Applying deep learning in e-commerce search. *arXiv preprint arXiv:2103.12982*.

A Appendix

A.1 Ranking Queries for Feature Training

During the training of IntenRank, we used specific queries for each semantic intensity feature to guide the model's understanding of these concepts. Below are examples of queries used for each feature:

Repurchase • "Do customers plan to repurchase this product in the future?"

Durability • "What is the average lifespan of this product?"

Satisfaction • "What is the overall customer satisfaction rating for this product?"

Comparison • "What are the advantages of this product over competitors?"

Adaptability • "How well does this product adapt to different uses?"

These queries were used to train the model to recognize and rank text based on the specific semantic intensity features. The full set of queries for each feature is available in the project repository.