A Counterfactual Explanation Framework for Retrieval Models

Bhavik Chandna[†] **Procheta Sen**[♣]
University of California San Diego, USA[†]
University of Liverpool, United Kingdom.[♣]
bhavikchandna@gmail.com[†] procheta.sen@liverpool.ac.uk[♣]

Abstract

Explainability has become a crucial concern in modern machine learning and deep learning because it improves model transparency. Information retrieval (IR) is no exception. In the existing literature on explainable IR, the emphasis has predominantly been on explaining relevance with respect to a retrieval model. Typical questions include why a document is relevant to a query, why one document is more relevant than another, or why a particular set of documents is considered relevant for a query. However, limited attention has been given to understanding why a document is not favored by a retrieval model for a query (e.g., why it does not appear in the top-K results). To address this gap, we study the following question: which terms should be added to a document to improve its ranking? This, in turn, helps explain which missing words contribute to a document not being favored by a retrieval model for a given query. We use a counterfactual framework to address this problem. To the best of our knowledge, this is the first attempt to study this specific counterfactual setting, namely, identifying which missing words can affect a document's ranking. Our experiments demonstrate the effectiveness of the proposed approach for both statistical models (e.g., BM25) and deep-learning-based models (e.g., DRMM, DSSM, ColBERT, MonoT5). The code for our approach is available at https://github.com/ThePro Paradox/Explainable-IR.

1 Introduction

The requirement of transparency of Artificial Intelligence (AI) models has made explainability crucial, and this applies to Information Retrieval (IR) models as well (Anand et al., 2022). The target audience plays a significant role in achieving explainability for an IR model, as the units of explanation or questions may differ based on the end user. For instance, a healthcare specialist, who is a domain expert but not necessarily an IR specialist, might want to understand the reasons behind a ranked suggestion produced by a retrieval model in terms of words used (Singh and Anand, 2019). On the other hand, an IR practitioner may be more interested in understanding whether different IR axioms are followed by a retrieval model or not (Bondarenko et al., 2022).

This study focuses on the perspective of IR practitioners. More specifically, we introduce a counterfactual framework for retrieval models that addresses their needs. Existing work in explainable IR (ExIR) has addressed questions such as why a document is relevant to a query (Singh and Anand, 2019), why one document is more relevant than another for the same query (Penha et al., 2022), and why a list of documents is relevant to a query (Lyu and Anand, 2023). These approaches primarily explain why documents are relevant or why they are ranked comparatively.

However, questions such as which missing terms make a document unfavorable to a retrieval model (i.e., keep it outside the top-K) remain largely unexplored. Such explanations can help IR practitioners understand how a retrieval model may need to be modified. For example, if a retrieval model—especially a neural IR model (Rekabsaz and Schedl, 2020)—systematically fails to favor documents because they lack certain gender-specific terms, then the model may need to be debiased.

This limitation becomes critical in real-world settings — such as patent search, legal case retrieval, and clinical information access — users and IR engineers frequently need to understand not only why a document was retrieved, but also why a potentially relevant document failed to appear in the top-Kresults. Missing a relevant document can have legal, financial, or safety-critical implications. In such environments, stakeholders require per-document, contrastive explanations

Findings of the Association for Computational Linguistics: ACL 2026, pages 38477–38490

that specify what information was absent from the document and prevented it from being retrieved.

With the motivation described above, the fundamental research question which we address in our work is **RQ1:** '*Which terms, if added to a document, would improve its ranking under a given retrieval model?*'

We frame **RQ1** as a counterfactual problem. Similar to prior work on counterfactual explanations in AI (Kanamori et al., 2021; Van Looveren and Klaise, 2021), we seek explanations that also change the model output, i.e., improve the rank of a document in an IR system. Our experimental results show that on average, in 70% of cases the solution provided by the counterfactual setup improves a document's rank for a given query and ranking model.

Our Contributions The main contributions of this paper are as follows.

- Propose a novel model-agnostic counterfactual framework for retrieval models.
- Estimate a set of terms that can explain why a document does not appear within the top-K results for a given query and retrieval model.
- Provide a comprehensive analysis using existing state-of-the-art IR models.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 describes the counterfactual framework used in our work, Section 4 describes the experimental setup and Section 5 discusses results and ablation study. Section 6 concludes with this paper.

2 Related Work

Counterfactual Explanations The xAI field gained significant momentum with the development of the Local Interpretable Model-agnostic Explanations (LIME) method (Ribeiro et al., 2016), which offers a way to explain any classification model. While models like LIME explain why a model predicts a particular output, counterfactual explainers address the question of what changes in input features would be needed to alter the output. Counterfactual xAI was first brought into the limelight in early 2010s with seminal work of Pearl (2018). The study in Karimi et al. (2020) provided a practical framework named Model-Agnostic Counterfactual Explanations (MACE) for any model. Subsequently, a series of models Kanamori et al. (2021); Van Looveren and Klaise (2021); Parmentier and Vidal (2021); Carreira-Perpiñán and Hada (2021); Pawelczyk et al. (2022); Hamman et al. (2023) were proposed for counterfactual explanation based on different optimization frameworks. In our work, we use counterfactual Explanation framework proposed in (Mothilal et al., 2020) (explained in detail in Section 3).

Explainability in IR Explainability in IR models can be broadly categorized into four areas: a) Pointwise Explanation b) Pairwise Explanation c) Listwise Explanation and d) Generative Explanation.

Pointwise Explanations show the important features responsible for the relevance score predicted by a retrieval model for a query-document pair. Popular techniques include locally approximating the relevance scores predicted by the retrieval model using a regression model (Singh and Anand, 2019).

Pairwise Explanations predict why a particular document was favored by a ranking model compared to others. The work in (Xu et al., 2024) proposed a counterfactual explanation method to compare the ranking of a pair of documents with respect to a particular query.

Listwise Explanations focus on identifying the key features for a ranked list of documents and a query. Listwise explanations (Yu et al., 2022; Lyu and Anand, 2023) aim to capture a more global perspective compared to pointwise and pairwise explanations. The study in (Lyu and Anand, 2023) proposed an approach which combines the output of different explainers to capture the different aspects of relevance. The study in (Yu et al., 2022) trained a transformer model to generate explanation terms for a query and a ranked list of documents.

Generative Explanations (Singh and Anand, 2020; Lyu and Anand, 2023) generally leverage natural language processing to create new text content, like summaries or justifications, that directly address the user's query and information needs. Model-agnostic approaches (Singh and Anand, 2020) have been proposed to interpret the intent of the query as understood by a black box ranker.

From the above mentioned category of explanations in IR, we focus on pointwise explana-

tion in our research scope. In pointwise explanation, rather than explaining what are the words which are relevant in a document for a particular query we address the research question what are the words which are required to improve the ranking of the document with respect to a query.

Search Engine Optimization techniques (Egri and Bayrak, 2014; Erdmann et al., 2022) generally uses different features like commercial cost, links to optimize the performance of the search engine. A major difference of the work in (Egri and Bayrak, 2014; Erdmann et al., 2022) with our work is we only consider the words present in a document as a feature. Our objective is to improve the ranking of a particular document concerning a specific query and a retrieval model rather than improving the ranking of a document concerning any query belonging to a particular topic.

3 Counterfactual Framework for Information Retrieval (CFIR)

Problem Statement Let d represent a target document that does not appear in the top-K retrieved results of a query q and retrieval model M. The objective in CFIR is to identify a set of terms w_i which, when added to d, improve its ranking with respect to q and model M.

The above mentioned setup for CFIR is formally defined in Equation 1 where *CFIR*, employs a counterfactual document generator $c_k(f_{\{M,q\}}, d)$ which takes as input a classifier $f_{M,q}$ and the document d to construct a counterfactual document d' such that d' is likely to get a higher rank (within top-K) than d for model M and query q. The objective of $f_{\{M,q\}} : R^{|V|} \rightarrow \{0, 1\}$ (where V is the vocabulary, described in detail in Section 3.1) is to predict given a query q and a retrieval model M if a particular document d will be within top-Kor not. The counterfactual explanation is defined as the set of words present in d' but not in d (i.e. output of Equation 1).

$$\begin{aligned} CFIR(q, M, d) &= c_k(f_{\{M,q\}}, d) - d \\ &= d' - d = \cup_{i=1}^{m}\{w_i\} \end{aligned} \quad (1)$$

3.1 Building Classifier ($\mathbf{f_{\{M,q\}}}$)

Similar to existing xAI (Ribeiro et al., 2016) approaches, the classifier $f_{\{M,q\}}$ in our research scope essentially locally approximates the behavior of a retrieval model M, for a query q and a subset of documents retrieved for the query q. However, in contrast to the regression model in (Ribeiro et al., 2016), we build a binary classification model to predict whether a document d will be ranked within the top-K results or not for a specific query q and retrieval model M.

For each query q, we build a classifier which predicts whether a document will be retrived in top-Kor not with respect to a Model M. To build this classifier we take top K documents from the

In the classifier setup, the top-K documents for a query q and retrieval model M represent class 1 and any other document not belonging to this class represents class 0. Theoretically speaking, if a corpus had N number of documents, then there will be $N - K$ documents which should have class label 0 and $N - K$ is a very large number in general which can cause class imbalance issue.

To avoid this issue, we choose only K documents from the set $N - K$. Out of this K documents we use the x number of documents for which we want to generate the explanations and then we choose randomly selected $K - x$ documents from the N-K set. K serves as a predefined threshold, typically set to values such as 10, 20, or 30. For $f_{\{M,q\}}$, each document d is represented as a word term frequency based feature vector, denoted as d_{vec}.

Formally, **Feature Vector for Classifier $\mathbf{f_{\{M,q\}}}$** is represented as $d_{vec} = \{tf_1^d, tf_2^d \ldots, tf_{|V|}^d\}$ where tf_i^d represents the term frequency of the word w_i in d. Using all the words from all the documents retrieved for a query to construct the vocabulary set V can pose challenges. Consequently, we take the union of the most significant n words from each document d using a function named $Imp(d)$ (explained in detail in Section 4) to construct V. $V = \cup_{i=1}^{K}\{\cup_{j=1,w_j \in Imp(d_i)}^{n} w_j\}$. Appendix D depicts a step-by-step algorithm to construct the feature vector for the classifier and Figure 5 in Appendix D shows one sample feature vector for the classifier.

Counterfactual Document Generator $\mathbf{c_k(f_{M,q}, d)}$ in Equation 1 follows an approach similar to that of Mothilal et al. (2020). Specifically, $c_k(f_{M,q}, d)$ generates k candidate counterfactuals $c_1^{maxIter}, c_2^{maxIter}, \ldots, c_k^{maxIter}$ (where $maxIter$ is the maximum number of iterations upto which loss function is optimized) for each document d, from which we randomly select a single counterfactual (d' in Equation 1) that involves only insertion of new words without

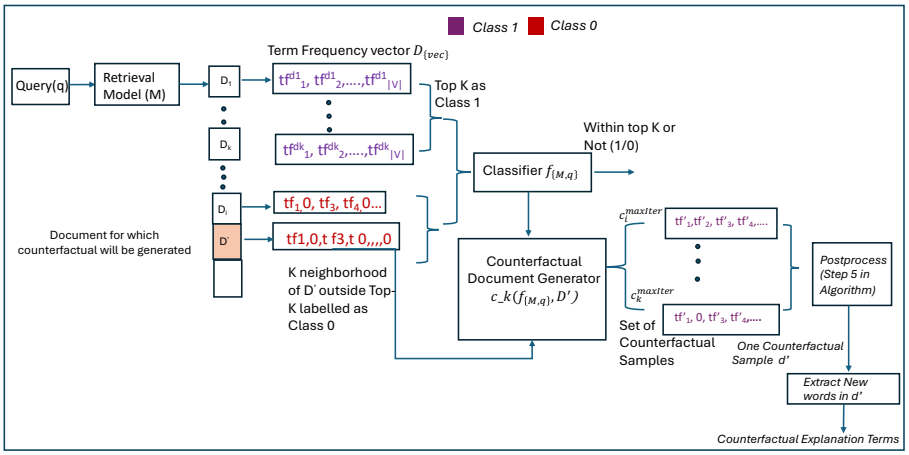


Figure 1: Schematic Diagram for Counterfactual Explanation Framework (CFIR)

modifying or deleting existing ones in d (step 5 in Algorithm 1). We fix k to a sufficiently large constant in our experiments. Similar to (Mothilal et al., 2020), the objective of $c_k(f_{M,q}, d)$ is to minimize three different criteria described as follows.

- **Criteria 1:** Minimizing the distance between the desired outcome y' (within top-K) and the prediction of the classifier model $f_{\{M,q\}}$ for a counterfactual example (c_i).
- **Criteria 2:** Minimizing the distance between any generated counterfactual (c_i) and the original document d. Broadly speaking, a counterfactual example closer to the original input should be more useful for a user.
- **Criteria 3:** Increasing diversity between generated counterfactuals.

Based on the above criteria, the loss function used to generate $c_1^{maxIter}, \ldots, c_k^{maxIter}$ is defined as follows.

$$\underset{c_1,\ldots c_k}{\arg\min} \left(\frac{1}{k} \sum_{i=1}^{k} \text{yloss}(f_{M,q}(c_i), y') + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, d) - \lambda_2 \text{div}(c_1, \ldots, c_k) \right) \quad (2)$$

In Equation 2, $yloss(\cdot)$ captures **Criteria 1**, $dist(c_i, d)$ captures **Criteria 2**, and div captures **Criteria 3**. The hyperparameters λ_1 and λ_2 in Equation 2 balance the contributions of the second and third terms of the loss function, i.e., similarity and diversity. Detailed definitions of the $yloss$, $dist$, and div functions used in Equation 2 are provided in Equations 4, 5, and 6, respectively, in Appendix G. We optimize the objective in Equation 2 using gradient descent.

Algorithm 1 shows step by step execution of the counterfactual document generator $c_k(f_{\{M,q\}}, d)$. In Algorithm 1 we show how the counterfactual examples $(c_1, \ldots c_k)$ are randomly initialized. The generated counterfactual examples (i.e. $c_i^{maxIter}$s) should change the prediction of classifier $f_{\{M,q\}}$ from 0 to 1 (i.e. modified document should be within top K). The set of words corresponding to the counterfactual explanation of d are the new words that have been added to d'_{vec} (i.e. feature vector representation of d' in Equation 1) compared to d_{vec}. Figure 1 shows the schematic diagram for counterfactual setup with the workflow between the different components (i.e. classifier and counterfactual document generator) within it.

Algorithm 1: CF Document Generator $c_k(f_{\{M,q\}}, d)$

Input : Classifier function: $f_{\{M,q\}}$, Feature Vector: $d_{vec} = \{tf_1, tf_2, \ldots, tf_{|V|}\}$, Number of Counterfactuals:k
Output : $\{d'_{vec} \in R^{|V|}\}$
Initialization:
  for $i \leftarrow 1$ **to** k **do**
    for $j \leftarrow 1$ **to** $|V|$ **do**
      $c^0_{i,j} = r \sim Random(.)$
      /* $c^0_{i,j}$ is the j^{th} coordinate of c_i at 0^{th} iteration */
    end for
  end for
for $t \leftarrow 0$ **to** $maxIter$ **do**
  Compute the loss $\frac{1}{k}\sum_{i=1}^{k} \text{yloss}(f_{M,q}(c_i^t), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(c_i^t, d) - \lambda_2 div(c_1^t, \ldots, c_k^t))$
  Update c_i^t's using gradient descent
end for
return d'_{vec}, d'_{vec} is a $|V|$ dimensional vector randomly chosen from the subset of $c_i^{maxIter}$'s for which $c_{i,j}^{maxIter} \geq tf_j^d \; \forall j = 1, \ldots, |V|$

4 Experiment Setup

Dataset We use three ranking datasets for our experiments: MS MARCO passage dataset for

passage ranking (Bajaj et al., 2016) and MS MARCO document ranking dataset for longer documents (Craswell et al., 2023) and TREC Robust (Voorhees, 2005) dataset. The MS MARCO passage and document ranking datasets contain queries from Bing[1] and the queries of TREC Robust are manually chosen. For each dataset, we randomly selected 100 queries from the test set and chose 5 documents not ranked in the top 10 results for each query, resulting in a test set of 500 query-document pairs. The details of the dataset are given in Table 3 in Appendix C.

We use five different retrieval models BM25, DRMM Guo et al. (2016), DSSM (Huang et al., 2013), ColBERT Khattab and Zaharia (2020), MonoT5 (Nogueira et al., 2020) and Splade (Formal et al., 2021) in our experiments. The details of each retrieval model is given in Appendix A.

Baselines To the best of our knowledge, this is the first work which attempts to provide counterfactual explanations in IR. Consequently, there exists no baseline for our proposed approach. However we have used a query word and top-Kword based intuitive baseline to compare with our proposed approach. In query word baseline (QW), we use query words not originally present in a document to enhance its ranking. For Top-K' ($Top-K'$) baseline we use the top k' words extracted from top 5 documents corresponding to a query as relevance set. Words appearing in the relevance set but not appearing in a document are added to the document to improve its ranking. For different retrieval models we have corresponding versions of QW and $Top-K'$ baselines.

Evaluation Metrics There exists no standard evaluation framework for exIR approaches. The three different evaluation metrics in our experiment setup are described as follows.

Fidelity (FD): Existing xAI approaches in IR use Fidelity (Anand et al., 2022) as one of the metrics to evaluate the effectiveness of the proposed explainability approach. Intuitively speaking, Fidelity measures the correctness of the features obtained from a xAI approach. In the context of the CFIR setup described in this work, we define this fidelity score as the number of times the words predicted by the counterfactual algorithm could actually improve the rank of a document. Let n be total number of query document pairs in our test case and x be number of query document pairs for which the the rank of the document improved after adding the counterfactuals obtained from the optimization setup described in Equation 2. Then the Fidelity score is mathematically defined with respect to a test dataset D and retrieval model M is defined as follows.

$$FD(D, M) = \frac{x}{n} * 100 \tag{3}$$

Avg. New Words: Here we compute the average number of new words added by the counterfactual approach for a set of query document pairs.

Avg. Query Overlap: Here we report on an average how many of the words suggested by the counterfactual algorithm come from the query words.

Parameters and Implementation Details The details of implementation about retrieval models are shown in Appendix B. We employed two popular classical machine learning methods, Logistic Regression (LR) and Random Forest (RF) for the classifier described in Section 3.1. For Logistic Regression, the learning rate was set to 0.001. For Random Forest, the number of estimators was set to 100. As described in Section 3.1, all the words present in a document are not used as input to the classifier. We use the top 10 ($n' = 10$) most important words from a document. As described in Section 3.1, we explored three different ways to implement $Imp(d)$ function a) TF-IDF weight based word extraction, b) BERT based keyword extraction (Grootendorst, 2020) and c) Similarity between the BERT representation of query and the document tokens. We found that BERT representation-based similarity computation worked the best for our approach. More details on the implementation of $Imp(d)$ function are shown in Appendix M. The value of K' for $Top-K'$ baseline was set to 5. More details on the parameter configuration are shown in Appendix H.

5 Results

Table 1 shows the performance of the counterfactual approach across different retrieval models (i.e., BM25, DRMM, DSSM, ColBERT, MonoT5, and Splade). We conducted experiments on the MS MARCO passage dataset, the MS MARCO document dataset, and the TREC Robust dataset to evaluate the effectiveness of our explanation

[1] https://bing.com

Model Description		MS MARCO Passage			MS MARCO Document			Trec Robust		
Retrieval Model	**Classifier**	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**
QW_{BM25}	NA	50%	5.61	100%	48%	6.14	100%	56%	6.12	100%
$Top-K'_{BM25}$	NA	42%	11.28	100%	40%	9.61	100%	41%	12.34	100%
$CFIR_{BM25}$	RF	65%	10.64	66%	52%	**16.81**	56%	**64%**	11.12	57%
$CFIR_{BM25}$	LR	**69%**	**17.14**	58%	**57%**	14.15	56%	58%	**13.25**	56%
QW_{DRMM}	NA	48%	5.12	100%	47%	6.14	100%	49%	7.12	100%
$Top-K'_{DRMM}$	NA	42%	**15.11**	100%	31%	14.12	100%	33%	**16.12**	100%
$CFIR_{DRMM}$	RF	**72%**	11.31	48%	56%	8.12	46%	62%	12.56	47%
$CFIR_{DRMM}$	LR	68%	12.37	62%	**62%**	**14.53**	45%	**65%**	13.47	43%
QW_{DSMM}	NA	49%	5.32	100%	45%	6.64	100%	52%	7.12	100%
$Top-K'_{DSSM}$	NA	35%	12.51	100%	32%	12.62	100%	34%	13.14	100%
$CFIR_{DSSM}$	RF	57%	11.52	58%	46%	18.14	57%	**59%**	12.46	100%
$CFIR_{DSSM}$	LR	**62%**	**15.78**	54%	**53%**	**18.52**	63%	58%	**17.24**	64%
$QW_{ColBERT}$	NA	56%	4.78	100%	34%	5.64	100%	38%	6.14	100%
$Top-K'_{ColBERT}$	NA	48%	**15.63**	100%	36%	**13.42**	100%	38%	11.32	100%
$CFIR_{ColBERT}$	RF	72%	12.41	56%	**72%**	11.05	49%	71%	10.35	52%
$CFIR_{ColBERT}$	LR	**75%**	14.12	61%	71%	10.23	62%	**74%**	**16.45**	65%
QW_{MonoT5}	NA	52%	10.15	100%	54%	12.23	100%	63%	10.15	100%
$Top-K'_{MonoT5}$	NA	75%	**14.11**	100%	68%	10.13	100%	75%	11.12	100%
$CFIR_{MonoT5}$	RF	80%	12.13	64%	72%	11.23	61%	73%	10.95	66%
$CFIR_{MonoT5}$	LR	**82%**	13.15	65%	**74%**	**12.23**	63%	**75%**	**11.45**	68%
QW_{Splade}	NA	49%	10.15	100%	51%	11.51	100%	62%	11.11	100%
$Top-K'_{Splade}$	NA	71%	**13.05**	100%	65%	9.23	100%	74%	12.22	100%
$CFIR_{Splade}$	RF	78%	11.23	62%	69%	12.11	60%	71%	9.81	65%
$CFIR_{Splade}$	LR	**80%**	12.15	63%	**71%**	**14.11**	64%	**73%**	**10.55**	67%

Table 1: CFIR model Performance for BM25, DRMM, DSSM, ColBERT, MonoT5 and Splade in MSMARCO Passage and Document Collection and TREC Robust. The Best Performing Counterfactual Explanation Method for every retrieval model is boldfaced; the overall best performance across all rows is underlined. All the results reported in Table 1 are statistically significant with $p < 0.05$.

approach across different document types. Four main observations can be drawn from Table 1. ***First***, the CFIR model for each retrieval model performs better than its corresponding query-word baseline or top-K words baseline in terms of fidelity score (FD). This observation is consistent across both passages and longer documents, i.e., MS MARCO passages, MS MARCO documents, and TREC Robust. ***Secondly***, it can be observed from Table 1 that mostly CFIR approach provided the highest number of new terms (terms not already present in the documents) as part of the explanation to improve ranking. Consequently, we can say the overall set of explanation terms are more diverse for CFIR approach compared to others. It can also be also observed from Table 1 that the Fidelity scores are generally better in the MS MARCO passages compared to MSMARCO document and TREC Robust dataset. One likely explanation for this phenomenon is that documents in MSMARCO document and TREC Robust are longer in length compared to passages. Consequently, it is easier for shorter documents to change the ranking compared to longer documents. ***Thirdly***, another interesting observation from Table 1 is that the maximum query word overlap by our proposed approach is 68%. This implies that the counterfactual algorithm is suggesting new words that are not even present in

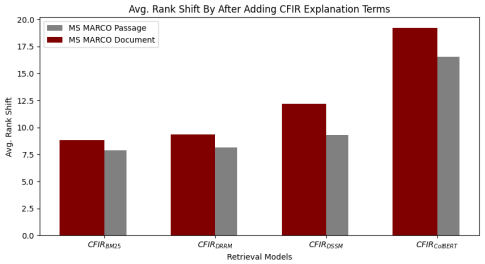


Figure 2: Average Rank shift by CFIR for BM25, DRMM, DSSM, ColBERT, MonoT5 and Splade

a query. ***Fourthly***, the performance of representation learning based retrieval models (i.e. ColBERT, MonoT5) are significantly better than the other models for Fidelity metric. One potential reason can be that, the counterfactual generator suggests words which are similar to the content of the document. Because of using better embedding representation (BERT (Devlin et al., 2019) and T5 compared to Word2Vec (Mikolov et al., 2013) in DRMM) these retrieval models give more priority to similar words than other retrieval models.

Prior work in information retrieval has explored adversarial attacks, where document content or embeddings are perturbed to manipulate rankings with malicious intent (Liu et al., 2023; Wu et al., 2022a). In contrast, the goal of counterfactual explanations is to provide interpretability for IR

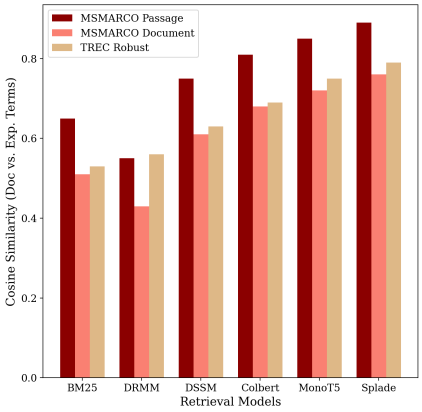


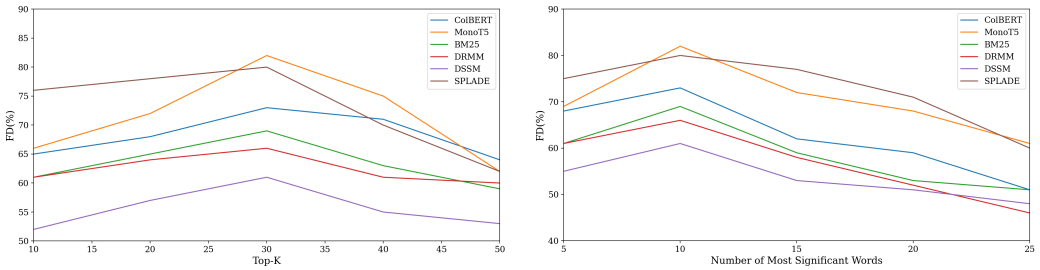
Figure 3: Average Semantic Similarity between original documents and the corresponding counterfactual explanation Terms for BM25, DRMM, DSSM, ColBERT, MonoT5 and Splade

models by revealing how document rankings can be improved. A key distinction lies in the nature of intervention: adversarial methods typically aim to introduce minimal perturbations often by substituting content, including important terms to preserve the original semantics while deceiving the model. In our case, CFIR explicitly seeks to identify new terms that, when added to a document, improve its rank, thereby highlighting what informative aspects were absent. Replacing important terms is not useful in counterfactual setup, as it fails to address what the document was lacking from the model's perspective. This formulation is particularly relevant for understanding model behavior, including uncovering potentially problematic model preferences (e.g., prior studies have observed gender bias in ranking systems). By identifying helpful additions, such as gendered terms, CFIR can reveal latent model sensitivities. Importantly, unlike adversarial attacks, the size of the added term set is also not constrained in CFIR (Avg. New Words column in Table 1 shows maximum 16.81 new words per explanation), as the focus is on explanatory sufficiency rather than minimality. However, for comparison, we have evaluated the performance of CFIR against the PRADA (Wu et al., 2022a) model which replaces certain words in a document to improve its ranking. Table 10 in Appendix L shows that CFIR performs better than PRADA for both ColBERT and MonoT5 in terms of Fidelity score. Table 7 in Appendix J shows a sample of example terms extracted by our proposed approach.

Further Analysis Figure 2 shows the average change in rank after introducing the explanation terms suggested by the CFIR setup. Figure 2 essentially demonstrates the actionability introduced by the counterfactual explanation terms. The two things to observe from Figure 2 are firstly, the average rank shift is greater for documents than for passages. Table 1 shows that ColBERT achieved a significantly higher fidelity score (16^{th} row) and a larger average rank shift compared to the other models, as also seen in Figure 2. Figure 3 shows the average cosine similarity computed between documents and the corresponding explanation terms. For both documents and the explanation terms we use pretrained BERT representations to compute the similarity. It can be observed from Figure 3 that the cosine similarity for the representation learning based retrieval models (i.e. ColBERT, MonoT5) are higher than the other retrieval models in general.

Parameter Sensitivity Analysis In Table 1, we observed that for most of the retrieval models the performance of the counterfactual explainer follows similar trend both in MSMARCO passage and document dataset (i.e. the best performing model in terms of fidelity score is same in most of the cases). As a result, we conducted parameter sensitivity experiments only on MSMARCO passage dataset. Figure 4 (a) shows the variance in Fidelity score with respect to the K value in Top-K. In Figure 4 (b) we show the variance of FD score with respect to the number of most significant words (i.e. n) used to construct the document vector. It is clearly visible from Figure 4 (b) that with an increase in the number of counterfactuals, there is a decrease in the performance of the counterfactual classifier. It can be observed that for $n = 10$ the best performance is achieved. Intuitively, as the number of words increases, the feature vector grows exponentially, making it challenging to train the classifier effectively.

Qualitative Evaluation of Explanations We conducted a user study involving three researchers with doctoral degrees in IR to estimate the quality of explanations. Each annotator was provided with 30 documents from the MS MARCO passage collection, along with the corresponding queries, ranked lists, and explanation terms generated by CFIR applied to the best-performing model, MonoT5 (shown in Table 1). Further details about the experiment setup is given in Appendix N. Users were asked to assess the quality



(a) (b)

Figure 4: Counterfactual Classifier Performance Variance with top-K and Counterfactual Performance Variance with variation of number of Counterfactuals

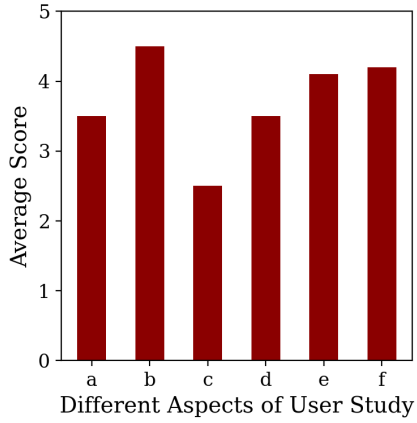


Figure 5: Qualitative Assessment of Generated Explanations over a) Intuitiveness b) Non-Intuitiveness c) Query Relatedness d) Document Relatedness e)Informativeness f) Diversity)

of explanations across six dimensions: (a) *Intuitiveness* how intuitive the explanation terms appeared given the query, document, and ranking context, with knowledge of the retrieval model; (b) *Non-intuitiveness* the extent to which explanations felt unexpected or misaligned with the query-document pair; (c) *Query Relatedness* whether the explanation terms were semantically related to the query; (d) *Document Relatedness* whether the explanation terms aligned with the overall topic of the document; (e) *Informativeness* whether the terms were meaningful and content-rich rather than generic or uninformative (e.g. mostly stop words); and (f) *Diversity* whether the explanation terms covered varied semantic aspects. For each aspect the users were asked to put a score between 0 to 5. Figure 5 shows that in general the explanation terms are intuitive and more similar to the document topic compared query topic (as expected due to use of document similarity criteria in the loss function in Equation 1). The explanation terms are also quite diverse. The non-intuitiveness score is quite low which shows that most of explanation terms follow an IR practitioner's intuition.

6 Conclusion

In this paper, we propose a counterfactual setup for a query-document pair and a retrieval model. Our experiments show that the proposed approach on an average 70% cases for both in short and long documents could successfully improve the ranking. In the future, we would like to explore different explanation units for the counterfactual setup.

7 Limitations

One of the limitations of this work is that we assume that top 10 or 20 words (based on tf-idf weights) within a document play the most important part in improving the rank of a document. However, theoretically speaking we should consider all the words present in a document to determine the most influential words for a retrieval model. We have used top tf-idf words (Similar to statistical retrieval models) to reduce the computational complexity of our experiments and we have seen that increasing the number of top words doesn't affect the performance of the model that much.

8 Ethical Considerations

In this work, we have used publicly available search query log and document collection to demonstrate counterfactual explanation. No sensitive data was used in this experiment. As a result of this there is no particular ethical concern associated with this work. If there is any kind of bias present in the search log data that effect can be observed within our approach. However mitigating that bias was beyond the scope of this work

References

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable information retrieval: A survey.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset. In *InCoCo@NIPS*.

Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic retrieval experimentation with ir_axioms. In *Proc. of SIGIR 2022*, pages 3131–3140.

Miguel Á Carreira-Perpiñán and Suryabhan Singh Hada. 2021. Counterfactual explanations for oblique decision trees: Exact, efficient algorithms. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6903–6911.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR*, abs/2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the trec 2022 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Gokhan Egri and Coskun Bayrak. 2014. The role of search engine optimization on keeping the user on the site. *Procedia Computer Science*, 36:335–342. Complex Adaptive Systems Philadelphia, PA November 3-5, 2014.

Anett Erdmann, Ramón Arilla, and José M. Ponzoa. 2022. Search engine optimization: The long-term strategy of keyword choice. *Journal of Business Research*, 144:650–662.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2288–2292.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2019. Matchzoo: A learning, practicing, and developing system for neural text matching. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1297–1300.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2023. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. 2021. Ordered counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11564–11574.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR.

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*, page 39–48.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2356–2362.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1647–1656, New York, NY, USA. Association for Computing Machinery.

Lijun Lyu and Avishek Anand. 2023. Listwise explanations for ranking models using multiple explainers. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 653–668, Berlin, Heidelberg. Springer-Verlag.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Axel Parmentier and Thibaut Vidal. 2021. Optimal counterfactual explanations in tree ensembles. In *International conference on machine learning*, pages 8422–8431. PMLR.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 4574–4594. PMLR.

Judea Pearl. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

Gustavo Penha, Eyal Krikon, and Vanessa Murdock. 2022. Pairwise review-based explanations for voice product search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 300–304.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2065–2068.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proc.of SIGKDD 2016*, page 1135–1144.

Joel Rorseth, Parke Godfrey, Lukasz Golab, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. 2023. Credence: Counterfactual explanations for document ranking. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, page 3631–3634. IEEE.

Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 770–773.

Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 618–628.

Arnaud Van Looveren and Janis Klaise. 2021. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer.

Ellen Voorhees. 2005. Overview of the trec 2004 robust retrieval track.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2022a. Prada: Practical black-box adversarial attacks against neural ranking models. *ArXiv preprint*, abs/2204.01321.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022b. Are neural ranking models robust? *ACM Trans. Inf. Syst.*, 41(2).

Zhichao Xu, Hemank Lamba, Qingyao Ai, Joel Tetreault, and Alex Jaimes. 2024. Counterfactual editing for search result explanation. *Preprint*, arXiv:2301.10389.

Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards explainable search results: a listwise explanation generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 669–680.

A Retrieval Models

The five different retrieval models used in our experiment are described as follows.

BM25: BM25[2] is a statistical retrieval model where the similarity between a query and a document is computed based on the term frequency of the query words present in the document, document frequency of the query words and also the document length.

DRMM: Deep Relevance Matching Model (DRMM) Guo et al. (2016) is a neural retrieval model where the semantic similarity between each pair of tokens corresponding to a query and a document is computed to estimate the final relevance score of a document.

[2] https://en.wikipedia.org/wiki/Okapi_BM25

DSSM: Deep Semantic Similarity Model (DSSM) Huang et al. (2013) is another neural retrieval model which uses word hashing techniques to compute the semantic similarity between a query and a document.

ColBERT: Contextualized Late Interaction over BERT (ColBERT) (Khattab and Zaharia, 2020), is an advanced neural retrieval model which exploits late interaction techniques based on BERT (Devlin et al., 2019) based representations of both query and document for retrieval.

MonoT5: MonoT5 (Nogueira et al., 2020) is a sequence-to-sequence model fine-tuned to predict the relevance of a query-document pair.

Splade: Splade (Formal et al., 2021)(Sparse Lexical and Expansion Model for Information Retrieval) combines the sparse interpretability of traditional IR models (like BM25) with the semantic power of deep learning. Unlike dense retrieval models that rely on vector similarity in embedding space, SPLADE encodes queries and documents into sparse high-dimensional vectors—essentially performing learned term expansions in a way that mimics the inverted index structure used in classic IR systems.

B Retrieval Performance of IR Models

We use Lin et al. (2021) toolkit for implementing BM25 and MonoT5 and Splade. For DRMM and DSSM, we use the implementation released by the study in Guo et al. (2019). For passage ranking we varied the parameters in a grid search and we took the configuration producing best MRR@10 value on TREC DL (Craswell et al., 2021) test set. For both DRMM and DSSM experiments on MSMARCO data, the parameters were set as suggested in (Wu et al., 2022b). The MRR@10 values are reported in Table 2 in Appendix B. For DRMM and DSSM, we use randomly chosen $100K$ query pairs from the MSMARCO training dataset to train the model.

The machine used to run the counterfactual experiments on the retrieval models had 1 A100 GPU and 40 GB of memory.

C Dataset Statistics

The dataset statistics for all the experiments are given in Table 3

	MRR@10	
Model	**MSMARCO Passage**	**MSMARCO Document**
BM25	0.1874	0.2184
DRMM	0.1623	0.1168
DSSM	0.1320	0.1168
ColBERT	0.3481	0.3469
MonoT5	0.3904	0.3827
Splade	0.3813	0.3721

Table 2: Retrieval Model Performance on MSMARCO passage and document

		MS MARCO Passage	MS MARCO Document	TREC Robust
Query	Avg Length	5.9	6.9	7.18
Document	Avg Length	64.9	1134.2	150.12
Query	#Instances	100	100	100
Document	#Instances	500	500	500

Table 3: Dataset Details for Counterfactual Setup

D Example of Input and Output to Classifier

Given an input query, we employ a LuceneSearcher with MSMARCO Index to retrieve the top-Kdocuments. The feature vector construction process follows these steps:

For each document, we:

1. Extract the top n words based on their Imp(d) values
2. Construct a vocabulary V as the union of all top 10 words across documents
3. Note that $|V|$ typically falls in the range of 150-180 words

The feature vector for each document has dimension $|V|$, where each component represents the value from the Imp(d) of the corresponding word from the vocabulary. Formally:

$d_{vec} \in R^{|V|}$

Labels are assigned according to the following criterion:

$$\text{label} = \begin{cases} 1 & \text{for top } K \text{ documents} \\ 0 & \text{for remaining documents} \end{cases}$$

Example feature vectors and their corresponding counterfactuals generated using (Mothilal et al., 2020) are shown in Table 5. Since $|V| = 150$ in our experiments, Table 5 shows only the term frequencies of the words present in each document. For all other words, the term-frequency values are zero in d_{vec}.

Existing Explanation Methods	Word Overlap
PointWise Explanation (Singh and Anand, 2019)	21.46%
ListWise Explanation (Lyu and Anand, 2023)	9.57%

Table 4: Comparison of CFIR with Existing ExIR Approaches

E Scalability Issues

There can be concerns regarding the feasibility of training a classifier per document. To address this, we propose and evaluate an alternative and more efficient strategy in which a single classifier is trained per query, rather than per document. Concretely, for a given query, we train one classifier using: (i) all documents for which explanations are required (let this number be x); (ii) their nearest neighbors, which contribute to the non-relevant document set; and (iii) the top-Kretrieved documents. To balance the number of relevant and non-relevant training instances, we construct the non-relevant set with a total size of 2k, where we sample 2k/x nearest neighbors from each document for which explanations are generated.

The results of this per-query training strategy are reported in Table 6. As shown, this substantially faster approach achieves performance comparable to that reported in Table 2, where a separate classifier was trained for each document. This demonstrates that our method remains effective while significantly reducing the computational overhead, directly addressing the reviewers' feasibility concerns.

F Existing EXIR approaches vs. CFIR

The existing literature aims to explain the significance of a document, a set of documents, or a pair of documents through various explanation methods. Nonetheless, our proposed approach diverges fundamentally from prior work in that we seek to demonstrate how the absence or frequency of certain tokens impacts document relevance. In this section, we examine whether there is any intersection between the two sets of tokens described earlier.

Pointwise Explanation Approach As outlined in Section 2, existing pointwise explanation methods elucidate why a specific document aligns with a given query within a retrieval model. Similarly, our proposed approach operates on individual documents and queries, albeit with a distinct objective. Here, we analyze the overlap between the explanations generated by the pointwise explanation method and those derived from our model, as presented in Table 4. This comparison was conducted on 50 pairs of documents.

Listwise Explanation Approach In Section 2, it is explained that listwise explanations typically aim to demonstrate the relevance of a list of documents to a given query. In listwise setup, one set of explanation terms are extracted for a list of documents, a query, and a retrieval model. Conversely, in our approach, we generate distinct explanations for each query-word pair. Therefore, to compare listwise explanations with our method, we aggregate all individual explanations obtained for each document-query pair in the list to create a unified explanation set for the entire list corresponding to a query. The resulting overlap is presented in Table 4.

G Counterfactual Optimization Framework

The components of Equation 2 are described here. The term $yloss$ in Equation 2 is a hinge loss, defined in Equation 4. In Equation 4, $z = -1$ when $y = 0$; otherwise, $z = 1$. The term $\text{logit}(f_{M,q}(c_i))$ denotes the logit value produced by the classifier $f_{M,q}$ when the counterfactual c_i is used as input.

$$yloss = \max\left(0, 1 - z\ \text{logit}(f_{M,q}(c_i))\right) \quad (4)$$

The distance function $dist(c_i, d)$ in Equation 2 is computed using Equation 5. In Equation 5, $|V|$ denotes the vocabulary size used to represent the document vectors (d_{vec}). The indicator function $I(c_p \neq d_p)$ is equal to 1 when the value of the p-th feature differs between the counterfactual input c and the original input d, and 0 otherwise.

$$dist(c, d) = \sum_{p=1}^{|V|} I(c_p \neq d_p) \quad (5)$$

The diversity term is defined in Equation 6. In this equation, the kernel entry $K_{i,j}$ is defined as $\frac{1}{1+dist(c_i,c_j)}$, where $dist(c_i, c_j)$ measures the distance between counterfactuals c_i and c_j.

$$div(c_1, \ldots, c_k) = \det(K) \quad (6)$$

H Parameters for Counterfactual Setup

The values of λ_1 and λ_2 in Equation 2 are set to 1 and 0.5, respectively. We set $k = 3$ in

docID	Feature Vector
3686955	[prohibition:2.0, amendment:2.0, under:1.0, dwindled:1.0, eighteenth:1.0, repeal:1.0, repealed:3.0, states:1.0, 1933: 1.0, ratification: 1.0]
6159679	[membrane:5.0, lipids:3.0, remainder:2.0, proteins:3.0, biochemical:2.0, 80:2.0, role:2.0, percent:2.0]
5217641	[waves:6.0, transverse:5.0, electromagnetic:3.0, oscillations:2.0, vibrations:2.0, travel:2.0, radiation:2.0, angles:2.0, transfer:2.0, types:3.0]

Table 5: Sample Feature Vector Corresponding to three different documents

Model Description		MS MARCO Passage			MS MARCO Document			Trec Robust		
Retrieval Model	Classifier	FD(%)	Avg. New Words	Avg. Query Overlap	FD(%)	Avg. New Words	Avg. Query Overlap	FD(%)	Avg. New Words	Avg. Query Overlap
$CFIR_{MonoT5}$	LR	**81.16**%	12.45	63%	**73**%	**13.13**	63%	**74**%	**10.45**	67%
$CFIR_{Splade}$	RF	78%	11.23	62%	69%	12.11	60%	71%	9.81	65%
$CFIR_{Splade}$	LR	**76.92**%	12.15	63.4%	**68**%	**11.33**	64%	**70.11**%	**8.91**	67%

Table 6: CFIR model Performance for MonoT5 and Splade in MSMARCO Passage and Document Collection and TREC Robust. The Best Performing Counterfactual Explanation Method for every retrieval model is boldfaced; the overall best performance across all rows is underlined. All the results reported in Table 1 are statistically significant with $p < 0.05$.

Equation 2. In all experiments reported in Table 1, we observed that for $K \geq 3$, we could always find a counterfactual explanation for each query-document pair using only word insertions to achieve the desired counterfactual outcome.

I Comparison with Credence(Rorseth et al., 2023)

The counterfactual method proposed in CREDENCE(Rorseth et al., 2023) identifies sentences whose removal leads to a decrease in a document's rank. In contrast, our work focuses on the complementary setting: identifying interventions that can increase a document's rank. Second, CREDENCE relies on the intuition that removing any sentence containing query terms lowers the ranking score. This is closely related to our query-word baseline, in which we instead add missing query terms to the document to improve its rank. Third, CREDENCE employs a heuristic procedure in which each candidate explanation document is repeatedly reranked and then verified to determine whether it is a valid explanation. In contrast, we introduce an explicit objective function that directly guides the optimization, thereby avoiding the expensive reranking loop required by heuristic approaches. Following the reviewer's suggestion, we now also include a direct comparison between CREDENCE and our proposed approach. To make the methods comparable, we use a reverse version of the CREDENCE heuristic by adding sentences that contain query terms to the document to improve its rank. The results are shown below.

J Example of Counterfactuals Produced by CFIR Setup

The words shown in Table 7 have improved the ranking of a docID with respect to the queries shown.

K Classifier Accuracy

The accuracy of the counterfactual classifier built in our setup is reported in Table 9.

L Adversarial Attacks vs. Counterfactual Explanation

Here we compare the performance of our proposed counterfactual explanation approach with PRADA, an existing adversarial model (Wu et al., 2022a). We use the MS MARCO passage dataset as the target corpus. In this experiment, we use the same test set described in Table 3 and used in the first column of Table 1. Table 10 reports the results in terms of fidelity score.

M Implementation of Imp(d)

We explored three ways to compute the top n words from each document. Each one of them is described as follows.

TF-IDF Approach: In this approach we choose top n words from a document based on their TF-IDF weight.

KEYBERT Approach: In this approach we use the model proposed in (Grootendorst, 2020) to extract keywords from a string.

BERT-Based Similarity(BERTSim): In this approach we compute the similarity between the BERT-based representation of the query text and

Retrieval Model	Query Text	docId	Explanation Terms
DRMM	What law repealed prohibition ?	3686955	working, strict, Maine, 1929, law, resentment, New York City, Irish, immigrant, prohibition, repeal, fall, Portland, temperance, riot, visit
DSSM	What is the role of lipid in the cell?	6159679	phospholipid, fluidity, storage, triglyceride, fatty receptor
ColBERT	what type of wave is electromagnetic?	5217641	directly ,oscillations, medium, wave, properties, speed
MonoT5	what is a caret?	6338711	display, diamond, weight
Splade	which vitamins help heal bruises?	3465680	minerals, body, eat, cut

Table 7: CFIR explanation terms for DRMM, DSSM, ColBERT, MonoT5 and Splade in MS MARCO passage.

	MS MARCO Passage			MS MARCO Document			Trec Robust		
Model	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**	**FD(%)**	**Avg. New Words**	**Avg. Query Overlap**
$CREDENCE_{BM25}$	64%	8.32	63%	52%	11.51	55%	56%	12.21	54%
$CREDENCE_{DRMM}$	68%	9.11	59%	58%	10.32	42%	61%	13.52	42%
$CREDENCE_{DSSM}$	60%	14.12	57%	51%	16.51	60%	54%	8.51	61%
$CREDENCE_{ColBERT}$	71%	11.21	54%	69%	10.52	52%	68%	12.56	53%
$CREDENCE_{MonoT5}$	71%	11.23	62%	70%	14.12	53%	72%	10.91	64%
$CREDENCE_{Splade}$	75%	9.72	65%	70%	10.35	62%	68%	9.21	59%

Table 8: Comparison of CFIR with Existing Counterfactual Approach (Rorseth et al., 2023)

Method	Accuracy
$Classifier_{BM25}$	81%
$Classifier_{DRMM}$	85%
$Classifier_{DSSM}$	82%
$Classifier_{ColBERT}$	84%
$Classifier_{MonoT5}$	86%
$Classifier_{Splade}$	88%

Table 9: Accuracy of the classifier used for counterfactual explanation.

Retrieval Model	**FD in PRADA**	**FD in CFIR**
ColBERT	74%	75%
MonoT5	80%	82%

Table 10: Performance of CFIR vs. Adversarial Attack Model PRADA (Wu et al., 2022a)

each token of the document and then we sort all the tokens based on the similarity.

Table 11 shows the performance of the above three approaches on the MS MARCO passage dataset with the ColBERT retrieval model. We set $n = 10$ for the experiments shown in Table 11. From Table 11, we conclude that the BERT-based similarity approach performs best for the $Imp(d)$ function. Hence, for all results reported in Table 1, we use the BERTSim approach for $Imp(d)$.

N User Study

In the user study, we did not record any personal information about the participants. We only recorded their judgments about the outputs of the proposed methodology. We also used data that are publicly available for IR research. Hence, no ethics approval was required for the study. All participants were informed that their assessments would be used in this research.

$Imp(d)$ **Approach**	FD
TFIDF	74%
KeyBERT	70%
BERTSim	**75%**

Table 11: Performance of Different Approaches in $Imp(d)$.