



Leveraging Out-of-the-Box Retrieval Models to Improve Mental Health Support

Theo Rummer-Downing^{1,2}^a and Julie Weeds¹^b

¹University of Sussex, Brighton, BN1 9RH, U.K.

²Tellmi, London SE1 7LL, U.K.

<https://www.tellmi.help/>

Keywords: Mental Health, BM25, SBERT, Information Retrieval, Semantic Search.

Abstract: This work compares the performance of several information retrieval (IR) models in the search for relevant mental health documents based on relevance to forum post queries from a fully-moderated online mental health service. Three different architectures are assessed: a sparse lexical model, BM25, is used as a baseline, alongside two neural SBERT-based architectures - the bi-encoder and the cross-encoder. We highlight the credibility of using pretrained language models (PLMs) out-of-the-box, without an additional fine-tuning stage, to achieve high retrieval quality across a limited set of resources. Error analysis of the ranking results suggested PLMs make errors on documents which contain so called *red-herrings* - words which are semantically related but irrelevant to the query - whereas human judgements were found to suffer when queries are vague and present no clear information need. Further, we show that bias towards an author's writing style within a PLM affects retrieval quality and, therefore, can impact on the success of mental health support if left unaddressed.

1 INTRODUCTION


The ability to find relevant information is now a simple and quick task thanks to algorithms like Google's PageRank (Brin and Page, 1998) and Okapi BM25 (Robertson et al., 1995). Information search usually involves submission of a *query* to a system that can return the most relevant results. However, the unifying assumption on which these systems are based is that a user has crafted a query in order to maximise the probability of retrieving relevant results. Although this paradigm suits searching for specific information, situations exist where information is sought indirectly, as is the case for the company, Tellmi.


Tellmi is a social enterprise which operates a youth mental health service via a mobile application. Tellmi was set up in order to tackle the mental health crisis facing young people. They cite a large number of statistics which indicate the poor state of youth mental health support, perhaps the most worrying of which being that suicide is the leading cause of death in young people (University of Manchester, 2018; Office for National Statistics, 2020). Tellmi service

users are supported using a fully-moderated peer-to-peer model, alongside access to a directory of mental health resources. The Tellmi directory contains over 500 useful resources which are available to users. Access to the resources currently requires either manual search by the user, or manual recommendation by an admin - both labour intensive and inefficient. It is proposed that having an accurate recommendation system for resources could significantly aid in signposting users to relevant help by reducing search times and increasing search accuracy. Further, accurate recommendation could enable Tellmi to scale the number of listings in their directory without increasing the difficulty for users to find relevant information.

Posts to the forum are typically statements, many of which lack a clear information request and hence, do not resemble *typical* IR queries which commonly display a 'direct' information need. For example, a user is unlikely to post "*How to tackle anxiety and worries?*". Instead, "*My worries are getting really bad and I am finding it hard not to give in to the worries*"¹ is more plausible in a forum setting. De-

¹It is important to note that any post written within this document has been paraphrased/re-written to protect anonymity.

^a <https://orcid.org/0000-0002-7052-0509>

^b <https://orcid.org/0000-0002-3831-4019>

spite this, it is possible that resources exist which could be highly useful and relevant to their problem, hence, an ‘indirect query’ problem whereby the information need is less explicit. The central inquiry of this work, therefore, is to determine whether existing relevance matching NLP algorithms can successfully retrieve relevant mental health resources based on the content of a user’s post. As an example, the NHS resource with the description “*Feeling low, anxious or worried? We offer a confidential space to help.*” should have high relevance to the aforementioned post, whereas the resource titled “*My Battle With Anorexia*” should be ranked with lower relevance to the query.

Many IR models rely on word embedding methods. Traditional methods are based on sparse embeddings produced by models like the Okapi BM25 algorithm. More recently, dense embedding methods have become the norm in NLP literature with a focus on deep transformer-based language models, such as BERT (Devlin et al., 2019), which are pre-trained on enormous corpora. The dense embeddings produced by these pre-trained language models (PLMs) have been successfully applied to IR methodologies (Reimers and Gurevych, 2019).

We focus on the implementation of two different SBERT (Reimers and Gurevych, 2019) architectures: the bi-encoder and the cross-encoder. These models use a pre-trained attention network to create contextualised vector representations so that relevance between documents can be calculated using a distance measure, such as the cosine similarity. The advantage these models present over statistical models is their potential to overcome instances which lack lexical overlap of query terms. For example, given a set of posts containing words such as *bulimia*, *binge* and *starve* it should be possible to match these with resources containing words like *eating disorder* and *overweight* due to their similar contexts, despite the resources containing no direct lexical overlap with the posts.

This paper contributes the following points to the health informatics literature. We show that pre-trained SBERT models can be applied out-of-the-box with high success on an information retrieval task within a health forum. However, the work also highlights the concept of stylistic bias within PLMs, and details the hindrance this bias has on the quality of retrieval. Further, we suggest that the ‘indirect query’ problem, whereby queries do not explicitly state an information need, is an under-addressed problem in information retrieval, and that current Question Answering (QA) datasets are likely to be insufficient for research into this specific problem as they model a

very ‘direct’ notion of relevance. More generally, we show the potential of using information retrieval techniques to improve access to health-related information.

2 RELATED WORK

The family of BM25 algorithms are some of the most successful text-retrieval algorithms that have been developed (Robertson and Zaragoza, 2009). The algorithm is a probabilistic method developed in the early 1990s for the Text Retrieval Conferences (TREC) based on term frequency statistics between a document and a query, and although somewhat outdated, it is still useful today as a high-recall method for retrieval prior to re-ranking via neural methods (Trablsi et al., 2021). The basis for all BM25 algorithms is the original Okapi BM25 algorithm detailed by Robertson et al. (1996) and used in the Okapi Information Retrieval system. Some variants which have been developed include BM25L, BM25+ and BM25-adpt, all of which are detailed by Trotman et al. (2014). Variants of BM25 are found to be a common baseline method across much IR literature.

Mikolov et al. (2013) introduced the Word2Vec model, which builds dense vector embeddings of words within a corpus of text by training a neural network architecture. These representations can be projected into a semantic space and compared with semantically similar words, and have proven useful for improving the generalisation of NLP models.

More recently, advanced neural models have utilised the transformer architectures described by Vaswani et al. (2017) and have become state-of-the-art on NLP benchmarks. Possibly the most well-known application of transformers is the work on Bidirectional Encoder Representations from Transformers, or BERT (Devlin et al., 2019). Devlin et al. (2019) describe BERT as a stack of encoder units² which is trained using a *masked language model* (MLM) pre-training objective. BERT set new state-of-the-art across eleven NLP benchmarks and, as a result, many variations based on this architecture have been developed (Liu et al., 2019; Sanh et al., 2020).

BERT-based models have been shown to have state-of-the-art performance on sentence-level tasks such as semantic textual similarity (STS Benchmark) (Devlin et al., 2019; Liu et al., 2019). However, Reimers and Gurevych (2019) describe how BERT would take approximately 65 hours to perform all-pairs similarity for 10,000 sentences, and they found

²Two BERT models were designed, *BERT_{BASE}*, with 12 encoder layers, and *BERT_{LARGE}*, with 24 encoder layers.

that the BERT sentence embeddings are inherently poor for comparing sentences in semantic space. Reimers and Gurevych (2019) proposed SentenceBERT (SBERT) to tackle this problem.

The SBERT model consists of two identical BERT models with tied weights. It relies on a pooling strategy over the outputs to create sentence embeddings of fixed-length. SBERT creates a more uniform semantic space for sentence comparison by ensuring the attention mechanism cannot attend to information from the comparative sentence. The results on the Semantic Textual Similarity (STS) benchmarks indicate that sentence embeddings produced by SBERT capture sentence-level semantics well, and are therefore well suited for sentence-level semantic similarity. Further, as SBERT computes embeddings before inference time, all-pairs similarity for 10,000 sentences takes approximately 5 seconds, making it commercially viable.

Knowledge distillation, whereby a small student model with relatively few parameters is trained to replicate the prediction distribution of a larger BERT teacher model, has been used by Sanh et al. (2020) to produce DistilBERT. In DistilBERT, the pre-training parameters of BERT are condensed into a more compressed model whilst retaining around 97% of BERT's performance. The resultant model leverages the full predictive distribution of BERT to enable high performance, whilst greatly reducing fine-tuning costs and increasing computation speed. Knowledge distillation was also used by Microsoft to create MiniLM (Wang et al., 2020), a smaller version of their UniLM (Dong et al., 2019) which is very similar in design to BERT. The student-teacher paradigm was also used in conjunction with newly designed loss functions to construct TinyBERT (Jiao et al., 2020), in which the authors successfully compressed the majority of the BERT teacher into only four encoder layers.

Knowledge-distilled models have successfully been used with the SBERT pooling layer for semantic search problems (Reimers, 2021b) by training on QA datasets. Specific *asymmetric*-search models use the MS MARCO Passage Ranking Dataset (Bajaj et al., 2018) due to its asymmetry between query and document lengths. On evaluation with a diverse range of semantic search tasks, Reimers (2021b) reported the highest performing bi-encoder to be a DistilBERT model, followed closely by a model based on MiniLM, although the MiniLM models outperformed DistilBERT in computation speed. The best cross-encoder was also based on MiniLM. Knowledge-distilled models are the main focus for this work.

Datasets for sentence embedding evaluation include the SentEval toolkit (Conneau and Kiela, 2018),

the GLUE benchmark (Wang et al., 2019), and the BEIR benchmark (Thakur et al., 2021). All contain various sentence-level tasks, with BEIR more directed towards information retrieval. However, the datasets for the retrieval tasks tend to use queries which have a direct relation to relevant documents, for example, the BEIR SciDocs dataset in which queries are abstracts of the larger articles.

There appears to be less research on IR tasks where links between queries and documents are less direct, hence the 'indirect' query problem appears to be an under-addressed problem in IR. Closely related problems tend to centre around forum threads, with tasks including retrieval of related threads (Cho et al., 2014; Elsas and Carbonell, 2009), duplicate query retrieval (Saha et al., 2019) and forum thread ranking (Faisal et al., 2016). Cho et al. (2014) describe how they also use a forum post as a query for a document retrieval task. However, most settings more closely resemble question-answer-type retrieval as they involve queries which are directly seeking information and documents (threads) which actively attempt to answer the query.

3 DATASET

The Tellmi dataset consists of text from user posts and the set of text descriptions of mental health resources. The content of the posts includes discussions about mental health, worries, complaints, and general concerns of young people, and are comprised of both statements and questions. The mental health resources vary greatly in length and style, and include: poems and stories written by both users and professional writers; advertisements for third-party organisations for self-improvement and learning; and descriptions for specific mental health charities and helplines which offer support. Summary statistics of the dataset are detailed in Table 1. As can be seen from the table, the dataset is highly asymmetric as the query length is considerably shorter than that of the document. Further asymmetry is characterised by the inability to logically reverse the search task as a resource does not tend to implicate a post - only the inverse is true. Although Tellmi provided a sample of over 180,000 posts, our evaluation focuses on a small number of posts in order to enable qualitative study.

Generally, the maximum input sequence length to BERT-like models is 512 tokens, with the default operation for sequences over the limit being truncation. With reference to Table 1, the posts are suitable for input to a BERT model. The lengths of resources have greater variation, with some resulting in truncation.

All data is suitable for input to BM25 as there is no maximum input sequence length.

Table 1: Dataset summary statistics.

Statistic	Posts	Resources
Total no. of docs	187,733	329
Mean doc. length (tokens)	40.0	173
Std. dev. doc. length	22.2	313
Min doc. length	1	15
Max doc. length	175	2540

4 METHOD

The methodologies employed in this work centre around the use of several architectures: lexical retrieval, and knowledge-distilled SBERT bi-encoders and cross-encoders. Specifically, we use BM25 for lexical retrieval, and we use the MiniLM SBERT model as a basis for the bi-encoder and cross-encoder. The design of an evaluation dataset for both quantitative and qualitative evaluation is described in section 4.1. Section 4.2 describes the BM25 algorithm, section 4.3 details the SBERT models which were used, and section 4.4 gives an overview of our experimental design.

4.1 Evaluation Dataset

The Tellmi dataset does not contain annotations for relevance between posts and resource documents. Therefore, it was necessary to build a small evaluation set in order to quantitatively and qualitatively evaluate the models. The original resource corpus was split evenly into development and test sets. We used the development set for preliminary work to qualitatively determine if it was possible to use a forum post as a query to retrieve relevant documents from the corpus using the retrieval methods detailed in sections 4.2 and 4.3. We then used the test set to create a quantitative evaluation set.

The dataset was built such that five manually pre-selected resource documents were ranked in order of their relevance to a post. The resources were manually chosen to ensure that, per query, three documents could be classed as relevant and two documents irrelevant, ensuring uniformity of ranking structure across different queries. Twenty posts were manually selected as queries, and the widely-used majority voting (Hernández-González et al., 2018) was applied across five annotators to produce a gold standard rank order of resources from 1 (most relevant) to 5 (least relevant) for each query. Later, one query was discarded due to low inter-annotator agreement

and subsequent inspection, leaving 19 queries in the evaluation set. To assess the reliability of the dataset, inter-annotator agreement was calculated using Krippendorff’s alpha (Krippendorff, 2019) from the Krippendorff python package (Castro, 2021). Following Krippendorff (2019) and Zapf et al. (2016), Krippendorff’s alpha was chosen as the agreement measure over Fleiss’ kappa (Fleiss, 1971) for several reasons: it is suitable for ordinal data; there is no upper bound on the number of annotators to compare; it is robust and allows for missing data; and it takes into account chance variation in ranking. Krippendorff (2019) describes how agreements with $\alpha \geq 0.800$ can be considered reliable, and below this threshold reliability tails off. We found the agreement across five annotators to be 0.814, indicating reliability of the dataset.

4.2 BM25

Our variant of BM25 is implemented using the ‘rank-bm25’ package (Brown, 2020). It uses a combination of the ATIRE BM25 function developed by Trotman et al. (2012) and the original Okapi BM25 algorithm detailed in Robertson et al. (1996). The model is described in equation 1:

$$rsv_{q,d} = \sum_{t \in q} IDF_t \cdot \frac{(k_1 + 1) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \quad (1)$$

where $rsv_{q,d}$ ³ is the calculated score for all terms, t , in the query, q , with document, d ; tf_{td} is the frequency of the term in the document; L_d is the length of the document; L_{avg} is the mean document length in the collection; k_1 and b are free tuning parameters; and the inverse document frequency function, IDF , is defined as:

$$IDF_t = \begin{cases} \log \frac{N - df_t + 0.5}{df_t + 0.5} & \text{if } IDF_t \geq 0 \\ \epsilon \cdot IDF_{avg} & \text{otherwise} \end{cases} \quad (2)$$

where N is the number of documents in the collection, df_t is the number of documents that the term appears in (the *document frequency*), IDF_{avg} is the mean IDF value over all terms in the corpus, and ϵ is a free tuning parameter.

The β , k_1 and ϵ parameters are important when considering the set up of BM25. We use grid search across the parameters to investigate an upper bound for model performance on the evaluation dataset. β controls the importance of the document length and k_1

³termed either “Retrieval Status Value” (Trotman et al., 2014) or “Robertson Selection Value” (Robertson et al., 1996), depending on the source

weights the term frequency, both of which operate on the term frequency section of the equation. ϵ is used as a scaling factor for the lower bound on the document frequency. When $k_1 = 0$, the query-document pair is scored solely on the occurrence of the term in the corpus and negates the contribution of the occurrence in the document. k_1 was varied between 0.01^4 and 3, at increments of 0.5. For $\beta = 0$, there is no penalisation for documents above average length. β was varied between 0 and 2, at increments of 0.25. $\epsilon = 0$ imparts no contribution from common words, and $\epsilon = 0.5$ weights common words by half of the average frequency across documents. ϵ was investigated between 0 and 0.5, with 0.1 increments. Intuitively, $\epsilon \geq 0.5$ is likely to greatly increase the noise in the system, hence ϵ was not investigated above 0.5.

Preprocessing for BM25 included tokenisation based on whitespace, lowercasing, punctuation removal, stop word removal and stemming. Stemming was performed using the NLTK implementation⁵ of the Snowball stemmer (Porter, 1980).

4.3 SBERT

Several bi- and cross-encoder models, which are available on the SBERT repository and listed in Tables 2 and 3, were selected for evaluation in this work based on their prior semantic search performance on six tasks from the BEIR dataset (Reimers, 2021b). These models have been fine-tuned on the MS MARCO Passage Ranking Dataset (Bajaj et al., 2018). Bi-encoders take a single sequence, in our case a post or resource, as input to produce a single embedding. Cross-encoders instead take two sequences simultaneously as input, separated by a separator token ([SEP]).

As with BM25, we aimed to investigate the upper bound for SBERT performance over the evaluation dataset by varying two parameters: the maximum sequence length of the input on both the bi- and cross-encoder, which alters the amount of data available for models to utilise in relevance measurements; and the input order of the query and document on the cross-encoder⁶, which has the potential to affect the relevance measurement as the position of the two sequences is encoded within the input. Sequence length was varied between 128 to 512 tokens at increments of 64, and sequence order was changed by placing the document before the query to form a document-[SEP]-query input.

⁴ $k_1 = 0$ can cause division by zero if a term is not present in the corpus, hence 0.01 was used as a lower bound.

⁵<https://www.nltk.org/api/nltk.stem.snowball.html>

⁶this is not a parameter of the bi-encoder

The only preprocessing to be performed for SBERT input was WordPiece tokenisation (Schuster and Nakajima, 2012). The MiniLM models only differ in their number of encoder layers (6 or 12), whereas the DistilBERT and TinyBERT models differ in their pre-training and architecture.

4.4 Experiments

4.4.1 Quantitative

The evaluation dataset was used to determine the upper bound on performance of the retrieval models. Mean Average Precision (MAP) was calculated across 19 queries, each with 5 respective candidate documents. MAP is described by Trotman et al. (2014) in equation 3:

$$MAP = \frac{\sum_{q \in Q} AP_q}{Q} \quad (3)$$

where q is a query in the query set, Q , and average precision, AP , is defined as:

$$AP_q = \frac{\sum_{n=1}^L \begin{cases} P_{qn} & L_n \text{ relevant} \\ 0 & \text{otherwise} \end{cases}}{N_{qr}} \quad (4)$$

where L is the number of documents in the results list; L_n is the relevance label at position n in the results list; and P_{qn} is the precision at position n , which is defined as the number of relevant documents in the result set up to document n , divided by n ; and N_{qr} is the number of relevant documents known for query q .

We use the MAP metric because it includes information about all relevant documents to a query and is suited to judging the rank order between binary relevance labels. We also give the standard deviation in the average precision, which indicates the variance across the different queries in the evaluation dataset. We chose not to use Mean Reciprocal Rank (MRR) as this only considers the first relevant document and discards useful information about subsequent relevant documents and, as there was a large proportion of relevant documents in the ranked candidates (3/5) compared with results from a much larger IR system, it was assumed that MAP would give greater resolution than MRR for results on this dataset. We decided against Normalised Discounted Cumulative Gain (NDCG) as this is best suited to judge the difference between relevance scores and, although this could be useful for future work, we are currently most interested in judgement of binary relevance labels.

A randomised baseline MAP score was bootstrapped across 1,000 iterations on randomly selected

data in order to determine a baseline performance on the dataset with which other models could be compared.

4.4.2 Qualitative

Qualitative evaluation and error analysis are used to determine and explain differences in quantitative performance on the evaluation set. Additionally, the models from Table 4 were used to retrieve the ten most relevant documents across the entire resource corpus (development and testing splits combined) for the 19 queries in the evaluation set, and the results were probed by one annotator to determine which models were found to best ‘answer’ the 19 queries. Reliability of the annotator was inferred from the high inter-annotator agreement on the evaluation dataset, and from their experience working for Tellmi.

4.4.3 Bias

We found that during curation of the evaluation dataset, two overarching writing styles within the resource corpus could be identified: *user-generated*, and *helpline-like* documents. The former were created by users of the app, and usually described their experiences with the majority written in the first person in an informal style. In contrast, the latter were written in either the second or third person and generally advertised a service.

We investigated the effect this could have on the aforementioned document retrieval task over the larger dataset. The proportions of each class of document in the top ten results were recorded for each model and compared.

Further, embeddings of the resource documents were created using a 6-layer MiniLM bi-encoder and clustered with the *community detection* clustering algorithm available in the Sentence Transformers package. Principal Component Analysis (PCA) was used to show the clusters in two dimensions in figure 1. The results of this investigation are described in Section 5.4.

5 RESULTS

Here we present and discuss results from the aforementioned experiments from Section 4.4. Section 5.1 describes the results of the ranking task experiments, followed by Section 5.2 where the model parameter experiments are discussed. The results of the qualitative evaluation are discussed in Section 5.3, and lastly, the results of the investigation into model bias are detailed in Section 5.4.

5.1 Ranking Experiments

Table 2: Bi-encoder model selection on the evaluation set (input sequence length: 512).

Bi-encoder Model	MAP \pm stdev.
msmarco-distilbert-cos-v5	0.91 \pm 0.12
msmarco-MiniLM-L6-cos-v5	0.89 \pm 0.13
msmarco-MiniLM-L12-cos-v5	0.91 \pm 0.13

Table 3: Cross-encoder model selection on evaluation set (input sequence length: 512).

Cross-encoder Model	MAP \pm stdev.
ms-marco-TinyBERT-L-2-v2	0.89 \pm 0.15
ms-marco-MiniLM-L-6-v2	0.96 \pm 0.08
ms-marco-MiniLM-L-12-v2	0.93 \pm 0.11

Table 2 compares MAP scores over the evaluation dataset for each bi-encoder model with a fixed input sequence length of 512 tokens, and table 3 presents the equivalent evaluation of cross-encoder models.

We find that the largest 12-layer bi-encoder model outperforms the smaller 6-layer bi-encoder models, whereas the 6-layer cross-encoder model returned the best performance of cross-encoders. We note there is no significant difference between the largest 12-layer bi-encoder and the DistilBERT bi-encoder, despite the latter being half the size of the former. Our results here are consistent with those of Reimers (2021b), who found that for these bi-encoder models there was no outstanding model across three separate datasets.

The 2-layer TinyBERT cross-encoder showed reduced performance compared to larger models which aligns with the results published by Reimers (2021a). However, we found that the 6-layer model outperformed its 12-layer counterpart which contrasts with Reimers (2021a) who found negligible difference between the two.

We present the best performing parameter settings for each retrieval algorithm over the evaluation dataset in Table 4. We observe that each model outperforms the random baseline, providing evidence that these methods are likely to be suitable for this type of retrieval task. We found the 6-layer MiniLM cross-encoder to have highest performance on the Tellmi dataset, with an MAP of 0.96 considerably close to the maximum achievable score of 1, indicating performance which was close to optimum for this task and, therefore, also close to expected human-level judgement.

With reference to Table 4, BM25 and the 12-layer MiniLM Bi-encoder were found to have comparable scores. This provided evidence against the hypothesis that, due to the complexity of the retrieval sce-

Table 4: Highest performance achieved by each method on the evaluation set.

Algorithm	MAP \pm stdev.	Parameters
Random Baseline	0.73 \pm 0.04	10,000 repeats
BM25	0.92 \pm 0.12	$k_1 = 1.5, \beta = 0, \epsilon = 0.4$
SBERT Bi-encoder	0.91 \pm 0.13	<i>max.len.</i> = 512, <i>model</i> = 12layerMiniLM
SBERT Cross-encoder	0.96 \pm 0.053	<i>max.len.</i> = 192, <i>model</i> = 6layerMiniLM

nario, both dense retrieval models would outperform the simpler lexical model. Quantitatively, it is evident that only the cross-encoder provides improved ranking over BM25.

The 6-layer cross-encoder was found able to outperform a bi-encoder of twice the layer size. This implies that the cross-encoder architecture requires less parameters to achieve equivalent results, and therefore, could be more applicable for relevance judgements. We found that the maximum input length had little effect on the results, implying that the majority of useful information is distributed towards the beginning of a resource.

Collectively, these observations indicate that it is theoretically possible to use these algorithms to rank relevant mental health resources by using a post as a query. The results on the dataset are somewhat limited in comparison with those over a typical ranking corpus due to the small dataset size. However, the results can be considered reliable due to the high inter-annotator agreement of the dataset noted in Section 4.1.

5.2 Parameter Evaluation

With reference to Section 4.2, we explored the effect of varying parameter settings for BM25. We found optima at $k_1 = 1.5$, $\beta = 0$, and $\epsilon < 0.4$. β and ϵ were found to have a slightly larger effect on performance than k_1 across the search space. The highest MAP score for BM25 on the evaluation dataset, using the aforementioned parameter settings, was 0.92 as shown in Table 4. We also note that across all parameter search values, MAP for BM25 varied between 0.88 and 0.92 and therefore consistently performed better than the random baseline MAP of 0.73.

Section 4.3 describes our exploration of the parameter settings for the SBERT models. Varying the maximum input sequence length to the bi- and cross-encoders had minimal effect on the MAP scores on the evaluation dataset, causing variance of only 0.02 to the respective models. We found that changing the input sequence order to the cross-encoder had a substantial detrimental effect on the MAP score, with a document-first order score of 0.78 compared with 0.96 for the standard query-first order - a reduction of 0.18. This was to be expected as the cross-encoders

were pre-trained on an asymmetric task and, as described in section 3, the Tellmi dataset is highly asymmetric. It follows that it is important to determine the symmetry of the task prior to use of the cross-encoder to ensure optimal performance.

We conclude this section by noting that completely out-of-the-box, i.e., without any parameter tuning, the bi-encoder and BM25 are likely to exhibit very similar performance. The MAP score for BM25 varied between 0.88 and 0.92 whereas the MAP score for the bi-encoder varied between 0.89 and 0.91. Assuming that the correct query-document input order is chosen, the MAP score for cross-encoder varied between 0.94 and 0.96, making it reliably the best out-of-the-box method.

5.3 Qualitative Evaluation

We interrogated the results of the ranking task based on the errors made by models and annotators. As three of the five documents for each query were relevant, we classed an error as any occurrence of a relevant document being ranked in the bottom two (or inversely, an irrelevant document occurring in the top three) compared with the majority-voted gold standard, as this indicated a fundamental error in relevance.

Upon qualitative study, we found that errors made by the cross-encoder mainly occurred where documents had relatively low lexical overlap with the query, but which included overlap of near-synonyms or related words which were irrelevant to the query - terms we denote as *red-herrings*. One example query contained '*suicide*'⁷ despite the topic being about grief. In this case, an irrelevant document which contained '*kill myself*', although being about bullying, was ranked highly. Hence, we suggest that the ability for dense models to match unseen synonymous and related words can be detrimental to performance as irrelevant documents can be assigned inflated similarity scores.

Conversely, we found that these errors were uncommon in the human annotation results. Instead, annotators were found to make errors where the in-

⁷These were large examples, hence only their important components are described.

formation need of a query was vague or ambiguous despite its topic being clearly defined. The queries were composed of statements and did not ask questions or explicitly seek advice. An example of this was “*My siblings don’t like me. Nobody likes me. I don’t speak to anybody and I’m really lonely*”. Interestingly, this was rarely a problem for the cross-encoder model, suggesting a strength of this model over human judgement.

As expected, the errors made by BM25, which resulted in some relevant documents being poorly scored, were because synonymous and related words could not be leveraged for relevance. However, this was relatively inconsequential to the quantitative results as the random baseline was still considerably lower, indicating that relevant documents within the dataset generally had some degree of lexical overlap with the query, and hence, BM25 could still be considered a useful model for a coarse or simple search option.

As described in section 4.4.2, document retrieval was performed over the entire resource corpus. We classed a query as ‘answered’ if one or more documents in the top ten results provided information to a user which was considered relevant and helpful. We found that the number of ‘answered’ queries were 12, 15 and 16 for BM25, the bi-encoder and the cross-encoder, respectively. These results suggest the dense retrieval models could be better suited to the task than BM25, and show that the dense models are more similar in performance over the larger dataset than the quantitative evaluation would suggest.

5.4 Bias Experiments

Upon investigation into potential bias within the resource corpus, as described in section 4.4.3, we found that the retrieval results from a bi-encoder were more likely to contain resources that were user-generated than helpline-like.

Table 5 describes the distribution of document types within the resource corpus, and Table 6 shows the average number of user-generated resources in the most relevant ($k = 10$) results returned by each retrieval method across 19 queries. We see that both dense retrieval models return a higher proportion of user-generated documents compared with BM25, indicating possible bias within the dense models towards this type. It is worth noting that although the number of helplines returned by the dense models were fewer, the number of *relevant* helplines was higher for these models, indicating higher precision.

As stated in section 4.4.3, clustering the first two principal components of bi-encoder document em-

beddings, as shown in Figure 1, was used to probe this further. Two relatively distinct clusters are observed which correlate strongly with their document type, indicating an observable difference between the two types. These clusters indicate retrieval bias towards the query *style* because similarity, and hence relevance, is affected by document style. Due to all queries being user-generated, the likelihood of a user-generated documents in search results is therefore increased.

Table 5: Proportions of resource types in corpus.

Resource Type	No. of Docs	% of Corpus
User-generated	44	13.4
Helpline-like	285	86.6
Total	329	-

Table 6: Average number of user-generated (UG) resources in top result set ($k = 10$) for each model.

Algorithm	Mean UG Docs. in Top 10 \pm stdev.
BM25	5.3 \pm 2.0
12-layer MiniLM Bi-encoder	6.1 \pm 2.1
6-layer MiniLM Cross-encoder	7.0 \pm 2.5

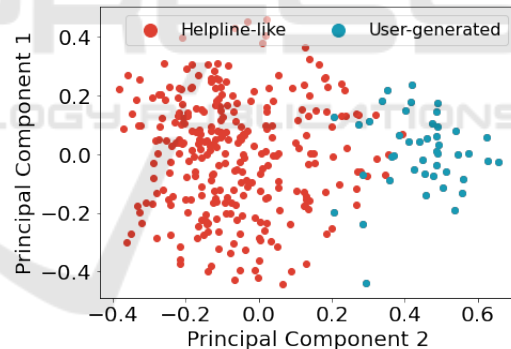


Figure 1: First two principal components from the 6-layer MiniLM bi-encoder resource embeddings.

6 CONCLUSIONS

In contrast to a typical search query where the query contains an explicit information need, we have considered the ‘indirect query’ problem whereby a search query does not explicitly request information, but to which documents may still exist that could be useful to the reader. We describe the example of a forum post on a mental health platform to which relevant mental health resources might be helpful. It has been shown that a forum post can contain sufficient information to obtain high quality results when used as a query on

a retrieval task. A traditional lexical method, BM25, was tested alongside two modern dense embedding methods, the bi-encoder and cross-encoder, for retrieval and ranking performance on a digital healthcare problem. All models performed above a random baseline on a small-scale ranking task, indicating that relevant documents can be ranked with a good level of success. Quantitatively, BM25 was equivalent in performance to a bi-encoder, however, a cross-encoder was found to be superior over both methods, with the 6-layer MiniLM model achieving the highest MAP score. Furthermore, a cross-encoder model of half the size of a bi-encoder showed equivalent performance, suggesting the cross-encoder architecture is more space efficient. Qualitative analysis over a larger corpus showed relevant documents returned within the top ten results, indicating good success with retrieval, however we also found that the dense retrieval models introduced bias towards the style of the query such that user-generated documents were more likely to occur in search results.

We suggest the following points as advice for practitioners within health informatics tackling similar information retrieval tasks. When maximum search performance is required, cross-encoders are likely to be the best solution. However, cross-encoders can only be deployed on small document sets because inference speed diminishes rapidly as corpus size increases, therefore bi-encoders should be the primary choice where any increase in scale is likely to be required. However, when high performance is not as critical, BM25 would be well suited to search at any scale, particularly because it has been found to perform above random without any changes to default parameter settings.

It is also important to understand the dataset that will be used in a search setting. In the case that the document and query sets are likely to be from similar distributions (and therefore unlikely to differ significantly stylistically), dense models are preferable to BM25 because the trade off between performance and bias would be favourable. In addition, when implementing a cross-encoder, understanding the symmetry of the search task will enable the practitioner to avoid using models which are fine-tuned on an incorrect task and which could severely degrade search performance. Symmetry can be judged by the degree to which document and query lengths differ, and by whether reversal of the search order would make a difference to the logic of the search. We recommend cross-encoder models trained on the MS MARCO dataset for asymmetric tasks. It is also suggested that the maximum sequence length parameter is unlikely to have a significant effect on performance, and there-

fore, dense models could be used without altering the default maximum.

For evaluation, we suggest that using MAP for binary relevance tasks is sufficient to show notable differences in performance, but that NDCG tends to be used for tasks involving relevance scores where the *degree* of relevance to a query is most important. Finally, we note that when using bi-encoder embeddings, dimensionality reduction via PCA and subsequent visualisation of embeddings can be a useful heuristic for bias evaluation.

The focus of this work was on ‘out-of-the-box’ neural models which have not had further masked-language model pre-training or fine-tuning for downstream tasks on in-domain data. Further work could therefore aim to evaluate the effect of both further pre-training and fine-tuning on retrieval and ranking performance. Development of a more robust evaluation dataset which encompasses a larger body of annotation work would improve reliability of findings. Investigation into methods to reduce bias within either the dataset or the models themselves could improve results and increase confidence in these methods.

ACKNOWLEDGEMENTS

This work is supported by Innovate UK. We also thank Colin Ashby for edits, the Tellmi annotators for their valuable time, and Jo Downing for her support.

REFERENCES

- Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. (2018). Ms marco: A human generated machine reading comprehension dataset.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117. Proceedings of the Seventh International World Wide Web Conference.
- Brown, D. (2020). Rank-BM25: A Collection of BM25 Algorithms in Python.
- Castro, S. (2021). Fast-Krippendorff: Fast computation of Krippendorff’s alpha agreement measure, based on Thomas Grill implementation.
- Cho, J., Sondhi, P., Zhai, C., and Schatz, B. (2014). Resolving healthcare forum posts via similar thread retrieval. *ACM BCB 2014 - 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 33–42.
- Conneau, A. and Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations.

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation.
- Elsas, J. and Carbonell, J. (2009). It pays to be picky: An evaluation of thread retrieval in online forums. pages 714–715.
- Faisal, M. S., Daud, A., Imran, F., and Rho, S. (2016). A novel framework for social web forums' thread ranking based on semantics and post quality features. *The Journal of Supercomputing*, 72:1–20.
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.
- Hernández-González, J., Inza, I., and Lozano, J. (2018). A note on the behavior of majority voting in multi-class domains with biased annotators. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1.
- Jiao, X., Yin, Y., Shang, L., Jiang, X., Xiao, C., Li, L., Wang, F., and Liu, Q. (2020). Tinybert: Distilling bert for natural language understanding.
- Krippendorff, K. (2019). *Content analysis : an introduction to its methodology*. SAGE, Los Angeles, fourth edition. edition.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pre-training approach.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.
- Office for National Statistics (2020). Leading causes of death, UK. publisher: GOV.uk.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 40:211–218.
- Reimers, N. (2021a). Ms marco cross-encoders — sentence-transformers documentation.
- Reimers, N. (2021b). Pretrained models — sentence-transformers documentation.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.
- Robertson, S., Walker, S., Hancock-Beaulieu, M. M., Gattford, M., and Payne, A. (1996). Okapi at trec-4. In *The Fourth Text REtrieval Conference (TREC-4)*, pages 73–96. Gaithersburg, MD: NIST.
- Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gattford, M. (1995). Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.
- Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Saha, S. K., Prakash, A., and Majumder, M. (2019). “similar query was answered earlier”: processing of patient authored text for retrieving relevant contents from health discussion forum. *Health Information Science and Systems*, 7.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
- Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.
- Trabelsi, M., Chen, Z., Davison, B. D., and Heflin, J. (2021). Neural ranking models for document retrieval.
- Trotman, A., Jia, X., and Crane, M. (2012). Towards an efficient and effective search engine. In *OSIR@ SIGIR*, pages 40–47.
- Trotman, A., Puurula, A., and Burgess, B. (2014). Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium, ADCS '14*, page 58–65, New York, NY, USA. Association for Computing Machinery.
- University of Manchester (2018). National Confidential Inquiry into Suicide and Safety in Mental Health.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.
- Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16.