ADOR: A New Medical Dataset for Sentiment-based IR

Mohammad Bahrani, Thomas Roelleke

Queen Mary University of London, UK

Abstract

Sentiment analysis has received attention in retrieval applications. Combining opinions such as user feelings with semantics would enhance the performance of these applications, especially when the level of urgency is essential, e.g., medical domain. However, no widely medical benchmark is known for evaluating sentiment-aware IR. In this paper, we create a dataset based on Amazon reviews for medical products and make it publicly available. To assess the compatibility of the benchmark with opinions and concepts we propose a sentiment-aware extension of TF.IDF and apply it to the dataset. This model is derived from linear combinations of sentiment-based TF.IDF score with term-based and conceptual TF.IDF scores. The benchmark could help healthcare organizations to effectively detect, rank and filter the most urgent notifications based on patient's health status, narratives and conditions.

Keywords

Semantic Retrieval, Query Analysis, Language Modelling, Benchmark, TREC, Query Formulation, Knowledge Representation,

1. Introduction

Despite the fact that both sentiment analysis and IR are of importance with regards to medical applications, the work on incorporating sentiments into medical IR is limited, and there is no well-known benchmark established for this task. Many review-based datasets have been released for the task of sentiment analysis such as multi-domain Amazon dataset [1], INEX social book search [2] and IMDB dataset of reviews [3]. However, researchers need a benchmark which primarily takes into consideration the integration of opinions and medical concepts. This is due to the importance of feelings in detecting the level of urgency in medical domain. Moreover, bio-medical companies need to analyse customer's general feelings about their products. On the other hand, patients need to know the sentiment of product reviews before buying. Wherefore the examination of sentiments would be beneficial for both buyers and suppliers of medical products.

In this paper, we address this problem by creating and making available a medical benchmark specifically for the task of opinion-aware retrieval.

Bio-medical benchmarks consider various pillars of semantics in collections and queries, e.g., terms, concepts and attributes. These semantics would enable data scientists to develop effective models for different tasks, e.g., filtering and classification.

Several benchmarks have been published to examine different IR models with respect to medical applications including OHSUMED [4], CLEF-eHealth [2, 5]. However, developing a sentiment-focused query-set for a dataset such as OHSUMED is not optimal since documents are generated from medical literature. Although sentiments, e.g., *cancer* and *treatment* are included in documents, implications of urgency and feelings e.g., emojis are rarely found. Table 1 shows the overview of well-known medical datasets which listed fundamental statistics of their semantic features.

Sentiment analysis and opinion mining are popular research fields in natural language processing, data science and text mining. They analyse textual contents based on people's opinions, emotions and attitudes [6]. In this paper, we create a benchmark that consists of a dataset, a query-set and the relevance results. The dataset consists of Amazon reviews for medical products. Additionally, it supports the use of common semantics (terms, concepts and relations) in biomedical retrieval.

The second contribution of this paper is to apply sentiment-aware models to the dataset. We propose a family of opinion-aware models for ranking medical reviews. These models are semantic instances of a generalizable TF.IDF. The technology of semantic retrieval is of particular importance in medical applications and the integration of semantics with the standard content-based retrieval tools could lead to more intelligent search experiences [7, 8]. The generalization of TF.IDF towards semantic frameworks is discussed in [9]. When compared to retrieval systems built upon only bag-of-words, the integrated methods result in more performant question answering (QA) systems with constraint checking abilities. There has been research on developing conceptual models for medical applications [10] and [11]. It could be interesting to leverage sentiments and feelings in these applications.

CIKM'21: Fourth Workshop on Knowledge-driven Analytics and Systems Impacting Human Quality of Life, November 01–05, 2021, CIKM, Australia

m.bahrani@qmul.ac.uk (M. Bahrani); t.roelleke@qmul.ac.uk (T. Roelleke)

CEUR Workshop Proceedings (CEUR-WS.org)

Dataset	Task	Reports	Number of Queries	*avg*-Opinions-per-query	*avg*-concepts-per-query
clef2013 e-health	Task3: Patients' Questions when Reading Clinical Reports	Overview of the ShARe CLEF eHealth Evaluation Lab 2013 [5]	50	0.3	2.9
clef2014 e-health	Task 3: use of information e.g. discharge summary and ontologies in IR	Overview of the share-clef ehealth evaluation lab 2014 [12]	50	0.34	1.86
OHSUMED	TREC-9 Filtering: Evaluate text filtering system	OHSUMED [4] - TREC-9 Final Report [13]	63	0.41	4.87
TREC 2006 Genomics Track	passage retrieval for Genomics question answering	TREC 2006 genomics track overview. [14]	27	0.32	6.00
TREC 2007 Genomics Track	Genomics passage retrieval based on biologists' needs	TREC 2007 genomics track overview.	35	0.27	4.6

Table 1: Overview of well established benchmarks for health-related retrieval.

By consolidating the methods for modelling opinions and sentiments in medical ranking, we aim to address the deficiencies in different tasks including but not limited to notification filtering and review filtering. In terms of notification filtering, we know that doctors and patients are overloaded with massive health-related data, and it is critical for health organizations to focus on the most important and urgent cases. In this scenario, the detection of urgency is associated with both ranking and acquisition of sentiments.

Our work contributes to building the grounds for improving medical review filtering through IR. It is the starting point of developing models that could better meet the needs of bio-medical organizations, companies and individual buyers for analysing most critical, positive and negative reviews.

2. The ADOR Dataset

The Amazon Dataset of Reviews (ADOR) is based on reviews from bio-medical Amazon products derived from three super categories which are Medication & Remedies, Diagnostic and Monitoring Tools and Health-Related Books. We have defined a set of sub-category products inherited from the super-categories and subsequently extracted reviews of related top ten items retrieved by Amazon search engine. However, in order to achieve a more balanced dataset in terms of polarity, we ignored items without negative reviews.

#Concepts	595442
#Distinct.Concepts	404748
#Opinions	194790
#Distinct.Opinions	163045
#Query	25
#Docs	44796
#Avg.Query Length	9.08
#Avg.Review.Text Length	35.38
#Sampling Date	31-03-2020

Table 2: The statistics of ADOR.

To make the data easily reusable, we followed two steps. Firstly, we converted the encoding of the contents to UTF-8 and secondly, we defined the schema and the required fields. The essential fields consists of Amazon ASIN number, medical category, star-rating, the title of the review, review text and labels including star-rating and helpful, have been embedded into the dataset.

2.1. ADOR Query Set

We have defined 25 topics based on five purposes. Figure 3 shows the distribution of queries and number of relevant documents. The five categories of information need are as follows:

1. The retrieval of positive or negative reviews associated with medical products.
2. Fact-based and non-sentiment-bearing queries which only intend to retrieve medical entities.
3. Ranking the polarity of item-reviews within the sub-categories, e.g. vitiligo cream and flu tablets.
4. Ranking the polarity of item-reviews within the super-categories, e.g. medications or diagnostic tools.
5. The retrieval of extreme (most positive or most negative) reviews given different medical concepts. We used modifiers to give attention to the information need, e.g. *Highly negative reviews for books about borderline personality disorder.*

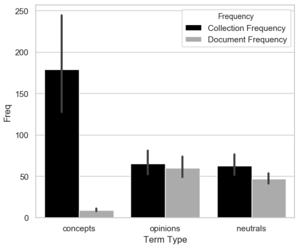

Figure 1: Document and collection statistics of the ADOR semantic types: The opinions group has the highest document frequency.

2.2. Overview of ADOR

In this section, we briefly present the dataset and provide the statistics of ADOR. Table 2 lists the fundamental statistics of the dataset. There are 194790 opinion features and 59442 medical concepts in the dataset which are distributed across 44796 documents. We used VADER lexicon to capture opinions and Meta-Map to bind terms to medical concepts. Figure 2 presents the distribution of document length and query length. The majority of queries (more than 35%) have a length between 9 and 12 words. More than 50% of documents have between 1 and 20 words, whereas 7% of them are longer than 100 words. The statistics regarding distribution of queries and their relevant documents are shown in Figure 3. As can be seen, 28% of queries contain 1-60 documents which is the exact same percentage for queries with more than 240 documents. The rest of the queries contain between 60 and 240 relevant documents. We extracted the average document and collection frequencies of semantic types (neutral terms, concepts and opinions) of the ADOR which can be found in Figure 1. Even though the average document frequency of opinions is high, opinions could significantly impact the retrieval quality due to the nature of reviews.

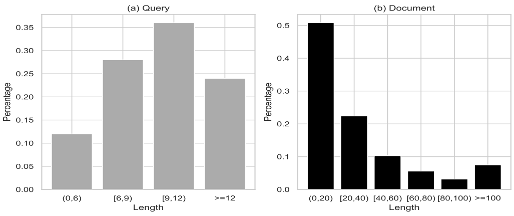

Figure 2: The distribution of document length and query length.

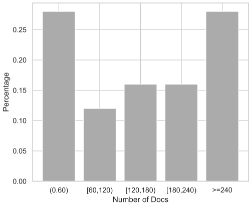

Figure 3: The distribution of queries and number of relevant documents.

3. Application of the Benchmark

3.1. Rationales

Although the use of human judgments could seem ideal for the generation of gold standards, we developed a generic framework which has some privileges, e.g., it could be easily used to build gold standards for new query sets.

We provided informative labels, including rating-star, the number of people who found reviews helpful and medical categories of Amazon products when preparing the data. This framework helps to rapidly develop new queries that could be formulated into the provided labels. Considering the example query *Why do some customers are happy with books about caffeine addiction and narcissistic personality disorder.*, the formulated query is : (Rating=[4,5], Super-Category=[Books], Sub-Category=[NPD,Caffeine Addiction]). In other words, any review in the dataset that meets the information needs requested by the formulated query could be selected.

To evaluate the accuracy of models, one approach would be the use of existing reviews as queries. However, there are two substantial issues with this approach. Firstly, data scientists need to analyse and classify their experimental results based on the query intent, e.g. fact-based, binary and explorative queries. The use of reviews as queries is not in line with the nature of query intent. Secondly, reviews are strongly focused on opinions. Therefore, generating a robust query set consists of a balanced combination of concepts, terms and opinions do interfere with the structure of reviews.

3.2. Baseline Models

The focus of this paper is to introduce a dataset for the task of semantic retrieval in the medical domain, sentiment-based and conceptual IR. Therefore, advanced ranking algorithms are the primary baselines. However, the benchmark is also able to be used for the prediction/classification tasks. For example, a review could be considered as a message posted by a patient or a customer. In this case, the evaluation approach is to predict if it is extreme (very negative) and requires attention by an expert, e.g., doctor, nurse or a company member. The other applicable task is notification systems. In this scenario, users post messages and an algorithm needs to decide who (e.g. which doctor, expert) should be notified for analysing the message or responding to it.

Furthermore, the framework could be employed by data scientists to predict features provided by the dataset such as positive/negative and helpful/not helpful. Baselines could be used such as Neural Network classifier (e.g., Bert or scikit), Bayesian predictor, regression and

Model	**Evaluation Measure**			
	P@5	**P@10**	**NDCG**	**MAP**
TF.IDF	0.2480	0.2720	0.2354	0.0833
BM25	0.3120	0.3160	0.2336	0.0813
KNRM	0.2320	0.2440	0.2445	0.0906
DSSM	0.2080	0.2200	0.2422	0.1039
arc-I	0.3520	0.3040	0.2476	0.0902
CF.IDF	0.3840	0.4080	0.2619	0.1106
OF.IDF	0.3680	0.4120	0.2758	0.1250
OF.IDF+TF.IDF (*w=0.5*)	0.3600	0.3920	0.2705	0.1175
OF.IDF+CF.IDF (*w=0.5*)	**0.4640**$^{\beta\theta\zeta}$	**0.4280**$^{\beta\theta\zeta}$	**0.2825**$^{\beta\theta\zeta}$	**0.1274**$^{\beta\theta}$

Table 3: Ranking performances of the opinion-aware models and the baseline methods: The bold font denotes the best result in that evaluation metric. β, θ, ζ indicate statistically significant improvements of the best model over **BM25**$^{\beta}$, **KNRM**$^{\theta}$ and **DSSM**$^{\zeta}$. The statistically significance is based on the paired t-test with p-value < 0.05.

K-NN (nearest neighbours) to measure the q prediction quality. The KNN classifier could be applied to retrieve the most similar train reviews (e.g., cosine similarity), aggregate evidence and assigns a label to the test review.

3.3. Processing the New Queries

To confirm the capability of the benchmark with models derived from opinions and concepts, we have developed a naive semantic approach. We briefly describe the methodology and then show the experimental results of comparing the semantic approach with well-known and recent IR methods on ADOR.

3.3.1. Methodology

Our approach is to leverage the well-known TF.IDF and capture its semantic extensions which are built upon opinions and/or concepts. To make the formulations readable, we use type-aware x functions, e.g. OF (o, d) is the opinion frequency of opinion o in document d, where CF (c, d) is the frequency of concept c in the document. Let q be a query, d be a document and let c be the collection, the Retrieval Status Value (RSV) of the opinion-aware model is as follows:

$$\mathrm{RSV}_{\mathrm{OF.IDF}}(d, q, c) := \sum_{o \in t} \mathrm{OF}(o, q) \cdot \mathrm{OF}(o, d) \cdot \mathrm{IDF}(o, c) \qquad (1)$$

IDF (o, c) is the Inverse Document Frequency of the opinion o in the collection. t is a list of all lexical features in lexicon where the sentiment polarity is equal to query polarity. For example, given query *Any useless or poor medications for allergy or cold sore.*, the query polarity is negative, and consequently, the t list comprises all negative opinions in the lexicon.

Let φ be a medical concept and let IDF (φ, c) be the Inverse Document Frequency weight of the concept, the conceptual extension of TF.IDF is defined as below:

$$\mathrm{RSV}_{\mathrm{CF.IDF}}\,(d, q, c) := \sum_{\varphi \in q} \mathrm{CF}\,(\varphi, q) \cdot \mathrm{CF}\,(\varphi, d) \cdot \mathrm{IDF}\,(\varphi, c) \qquad (2)$$

3.3.2. Evaluation

In this section, we briefly discuss the evaluation results of the propose semantic models, TF.IDF, BM25 and neural ranking models when applied to ADOR.

We have trained neural ranking models including KNRM [15], DSSM [16] and arc-I [17] on ADOR. We performed 5-fold cross-validation where the final fold in each run was considered as the test set. We randomly divided queries into five-folds and repeatedly captured the average of the fivefold-level evaluation results. All neural models were developed using MatchZoo [18] based on *tensorflow* with Adam optimizer, batch size 16 and learning rate=0.001. Using the Lucene framework and the Language Modelling with Dirichlet Prior, we retrieved pseudo-relevant documents and subsequently, the top 100 documents were re-ranked by the models. In addition to OF.IDF and CF.IDF, we conducted experiments on linear combinations of opinion-aware TF.IDF with term-based and conceptual TF.IDF using aggregation parameter $w = 0.5$. Concerning concept-based models, we used MetaMap to extract concepts accompanied by their frequencies, semantic types and scores. We counted 'trigger' attributes of MetaMap-outputs to calculate the corresponding frequencies of semantic types.

Table 3 shows the experimental results on ADOR using four metrics including P@5, p@10, NDCG and Mean Average Precision (MAP). We also conducted the paired t-test with $p < 0.05$ to compute the significance of improvements. The isolated OF.IDF and CF.IDF worked better than TF.IDF, BM25 and neural models (KNRM, DSSM, arc-I) while the combination of opinions and concepts received the best results. The interesting finding is that the models based on combinations of opinions with both terms (OF.IDF+TF.IDF) and concepts (OF.IDF+CF.IDF) improved all the measures.

4. Conclusion

In this paper, we introduced a new benchmark, namely ADOR which is a subset of Amazon reviews. For our research aim, the dataset allows for bringing and testing sentiment-based IR to medical domain. The corresponding dataset focuses on medical products within three categories including medicine, monitoring tools and health-related books. The collection of reviews comes with a structured framework which enables users to automatically generate relevance labels for new topics. Moreover, a query set with relevance results was consolidated into the benchmark. In order to develop this query set, we considered factors such as query intent, sentiment score of query and concept query frequency.

To measure the suitability of the benchmark for sentiment-based IR, we proposed naive but reproducible opinion-aware models as semantic instances of the generalizable TF.IDF. These models are derived from combinations of sentiment-only TF.IDF with term-only and concept-only TF.IDF. We compared the new approach with well-established and modern retrieval models. Our experiments confirmed that the integration of sentiments with IR improves the quality of ranking with regards to the ADOR dataset. The semantic model based on combination of OF.IDF and CF.IDF achieved the best results against gold standards.

In conclusion, the ADOR benchmark could help researchers to develop and evaluate opinion-aware retrieval models. These models would benefit companies and healthcare organizations to effectively detect, rank and filter urgent notifications based on patient's health status, narratives and conditions. The benchmark is available at https://github.com/mb320/ADOR.

References

[1] S. Li, C. Zong, Multi-domain sentiment classification, in: Proceedings of ACL-08: HLT, Short Papers, 2008, pp. 257–260.

[2] M. Hall, H. Huurdemann, M. Skov, D. Walsh, et al., Overview of the inex 2014 interactive social book search track, in: Conference & Labs of the Evaluation Forum (CLEF), 2014.

[3] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, Association for Computational Linguistics, 2011, pp. 142–150.

[4] W. Hersh, C. Buckley, T. Leone, D. Hickam, Ohsumed: an interactive retrieval evaluation and new large test collection for research, in: SIGIR'94, Springer, 1994, pp. 192–201.

[5] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al., Overview of the share/clef ehealth evaluation lab 2013, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013, pp. 212–231.

[6] B. Liu, Sentiment analysis and opinion mining, Synthesis lectures on human language technologies 5 (2012) 1–167.

[7] R. Van Zwol, T. Van Loosbroek, Effective use of semantic structure in xml retrieval, in: European Conference on Information Retrieval, Springer, 2007, pp. 621–628.

[8] M. Bahrani, T. Roelleke, FDCM: Towards balanced and generalizable concept-based models for effective medical ranking, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1957–1960.

[9] H. Azzam, S. Yahyaei, M. Bonzanini, T. Roelleke, A schema-driven approach for knowledge-oriented retrieval and query formulation, in: Proceedings of the Third International Workshop on Keyword Search on Structured Data, ACM, 2012, pp. 39–46.

[10] E. Meij, D. Trieschnigg, M. De Rijke, W. Kraaij, Conceptual language models for domain-specific retrieval, Information Processing & Management 46 (2010) 448–469.

[11] C. Wang, R. Akella, Concept-based relevance models for medical and semantic information retrieval, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 173–182.

[12] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, et al., Overview of the share/clef ehealth evaluation lab 2014, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2014, pp. 172–191.

[13] S. Robertson, D. A. Hull, The trec-9 filtering track final report, in: TREC, volume 10, Citeseer, 2000,

pp. 344250–344253.

[14] W. R. Hersh, A. M. Cohen, P. M. Roberts, H. K. Rekapalli, Trec 2006 genomics track overview., in: TREC, volume 7, 2006, pp. 500–274.

[15] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval, 2017, pp. 55–64.

[16] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 2333–2338.

[17] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Advances in neural information processing systems, 2014, pp. 2042–2050.

[18] Y. Fan, L. Pang, J. Hou, J. Guo, Y. Lan, X. Cheng, Matchzoo: A toolkit for deep text matching, arXiv preprint arXiv:1707.07270 (2017).