Weak Supervision for Question and Answering Sentiment Analysis Models

Victor A. K. Tomita[1] and Ricardo M. Marcacini[1]

Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP)
São Carlos - SP, Brazil
{akihito012,ricardo.marcacini}@usp.br

Abstract. The growth of social networks, e-commerce, and journalistic media has resulted in the proliferation of opinions on various topics. Companies and government agencies are interested in understanding their customers' opinions about their products and services. Automatic sentiment analysis methods can be used to extract the general sentiment about a product. However, traditional sentiment analysis methods are inflexible in dealing with human queries, which tend to ask questions. Therefore, question-and-answer (QA) systems for sentiment analysis offer a promising alternative. This paper proposes a new method called Weak Supervision for Question and Answering Sentiment Analysis (WSQASA) that fine-tunes and extracts sentiment through QA models in an unsupervised manner. We investigate question-generation models associated with sentiment filters for weak supervision, generating domain-specific question-and-answer pairs for fine-tuning the QA model. Our method enables the generation of domain-specific question-and-answer pairs for fine-tuning the QA model, which significantly enhances the QA-based sentiment analysis results, even without the usage of labeled data.

Keywords: Weak Supervision, Question Generation, Question and Answering, Sentiment Analysis

1 Introduction

The rapid growth of media and social networks, such as Twitter and Instagram, has resulted in enormous data containing opinions on a wide range of topics [22]. Various sectors, including companies and political parties, seek to leverage this data to gain insights into their audience's perceptions of their products and political candidates. For instance, companies aim to extract valuable information from reviews to improve product development and marketing strategies [12]. Similarly, political parties are interested in understanding the population's voting intentions by analyzing public opinions on political matters [18]. The vast amount of data generated by social media has become a crucial resource for various sectors, driving the need for efficient techniques to analyze this unstructured data.

Manual analysis of this data is impractical and time-consuming. To address this issue, sentiment analysis has been developed as a field aimed at automatically extracting emotional reactions expressed by users [29]. In particular, aspect-based sentiment analysis (ABSA) seeks to identify the sentiment expressed about a specific aspect of an object [27]. Typically, ABSA involves the extraction of entities accompanied by an aspect-based sentiment analyzer, which enables the extraction of key points about the product. However, this approach lacks flexibility in addressing user queries written in natural language. To overcome this limitation, research on question and answer (QA) systems for extracting textual snippets expressing sentiments have emerged [19]. In these systems, users explore different positive and negative aspects of products and services through questions, and the answers are snippets extracted from opinion texts.

Expanding data in specific scenarios, such as sentiment analysis through question and answer (QA) systems, does not necessarily correspond to an increase in labeled data. This issue is particularly problematic for supervised learning-based approaches. That said, weak supervision techniques have been introduced to leverage unsupervised data for model training [30]. This method allows the exploration of heuristics, filters, rules, or labeling functions to extract specific patterns in the data and automatically infer labels.

This paper proposes a weak-supervision method for training QA models for sentiment analysis to improve the model performance for domain-specific scenarios without labeled data. To achieve this goal, weak supervision techniques are explored through generative models to create synthetic data. In particular, we focus on BERT-based QA methods. We show that fine-tuned models using our weak-supervision method outperform general QA models in the sentiment analysis task, even without using labeled data. Our proposed method is called WSQASA (Weak Supervision for Question and Answering Sentiment Analysis), and our main contributions are :

- We present question generation (QG) models to generate labeled question and answer datasets oriented toward opinion words. These datasets are used within weak supervision to generate QA-based sentiment models without human-labeled data.
- We demonstrate the importance of domain-specific filtering, which allows for the training of QA systems for this kind of scenario. The proposed method seeks to extract aspects and simultaneously classify the polarity of sentiments for specific domains, and we report significant gains in this scenario.

2 Background and Related Works

Question-answering (QA) is a sophisticated information retrieval technique that employs natural language queries or statements to extract specific elements from a document, including texts, images, and videos. While such queries may precisely specify the desired information, they provide more extensive information than a simple list of search terms. The syntactic and semantic complexity of natural language defines the QA task [17].

The field of question-answering has witnessed significant progress with the advent of deep neural network (DNN) based methods [8]. This can be attributed to the constantly increasing size of data, which allows for effective fine-tuning of DNN weights. In contrast, traditional QA approaches rely on feature engineering methods augmented by machine learning algorithms, which require considerable effort in selecting or creating features [5]. However, deep learning (DL) facilitates end-to-end learning for QA tasks, alleviating the feature creation challenge [8]. DL-based machine learning methods can extract patterns present in data, which can be applied to various tasks, making DL methods highly versatile as they do not rely on rules specific to particular domains [3].

A significant feature of recent architectures is the attention mechanism of the transformer architecture. This mechanism addresses the problem of information loss during signal processing up to the decoder and determines the relationship between two parts of sequences [20]. Specifically, the attention mechanism is used to identify which words in sequence B should receive more attention when considering words in sequence A, making it well-suited for QA tasks. Notable models employed for QA tasks include the Generative Pre-trained Transformer (GPT) [4], which consists of a 12-layer decoder comprising only transformer structures with masked self-attention heads. Another model is the DistilBERT [26], which is trained via a distillation process, with BERT as the teacher model.

Transformers-based QA models have been recently used to support sentiment analysis. In their work, Mishra et al. (2022) [19] propose a QA model capable of extracting a snippet from a tweet to explain the sentiment associated with the text. To accomplish this, they trained BERT, DistilBERT, AlBERT, and RoBERTa, identifying DistilBERT as the best-performing model. Similarly, our work employs the DistilBERT model to demonstrate how our approach can enhance QA-based sentiment analysis even in scenarios without labeled data.

Table 1 illustrates an example of a sentiment analysis task through QA. The first column shows the context, while the second column displays the question. These inputs are used in the QA model to predict the answer in the last column.

Context	Question	Answer
Latest update randomly crashes and has major issues loading chunks	what are the issues?	has major issues loading chunks
Designed for professionals, and thus is too complicated	Is the design good ?	is too complicated
The battery is pretty good, but the performance is very bad	what are the problems with the product?	performance is very bad

Table 1. Example of the task of sentiment analysis through QA

One of the main challenges of training QA models is the requirement for large sets of accurately labeled data. However, this can be a time-consuming and expensive process. Weak supervision methods solve this issue by enabling models to be trained with limited labeled data. Weak supervision techniques rely on leveraging various sources of information, such as heuristics, rules, or distant supervision, to train models with incomplete or imprecise labels.

Despite its potential usefulness for many practical applications, there is a lack of literature on how to use weak supervision to refine QA models for sentiment analysis. In particular, there is a challenge in generating annotated data in which the answers to the questions are oriented toward opinion words. Existing works have mainly focused on using weak supervision for training QA models on large-scale datasets without considering the specific requirements for sentiment analysis [1]. Addressing this research gap can have important implications for real-world applications, as sentiment analysis is widely used in industries such as marketing, customer service, and political analysis. By leveraging weak supervision methods, it may be possible to improve the accuracy and efficiency of QA models for sentiment analysis while reducing the reliance on large sets of labeled data.

3 Method

We propose a method for QA-Based Sentiment Analysis that is trained under weak supervision. Our approach aims to predict the optimal answer span containing sentiment words from all possible candidates of a given document d, referred to as $A(d)$ [14]. Formally, we define the probability distribution over all feasible answer spans, given a question q and a passage p. The predictor function identifies the answer with the highest likelihood, as defined in Equation 1.

$$f(q, d) = \arg \max_{a \in A(d)} P(a|q, d) \tag{1}$$

Many approaches can be used to calculate the probability $P(a|q, d)$, which states the likelihood of the answer a given a question q and a document d. Our method uses DistilBERT to predict the start and end tokens associated with the answer a.

Our proposed method, WSQASA, addresses the challenge of limited labeled data in QA-based sentiment analysis. We adopt a three-step approach, as Figure 1 illustrates. Firstly, we leverage weak supervision techniques via a generative QA model to generate domain-specific training datasets. Secondly, we propose a sentiment filter based on a sentiment lexicon and sentiment classification models. This filter allows the QA model to extract snippets from contexts expressing sentiments. Finally, we fine-tune a pre-trained QA model on the filtered training data. This approach enables us to generate metrics that can be compared to the original pre-trained QA model, which was not fine-tuned.

3.1 Weak Supervision using Question Generation

To augment the quantity of data for model training, we utilize Question Generation (QG) methods [6]. QG is designed to generate factoid questions automat-

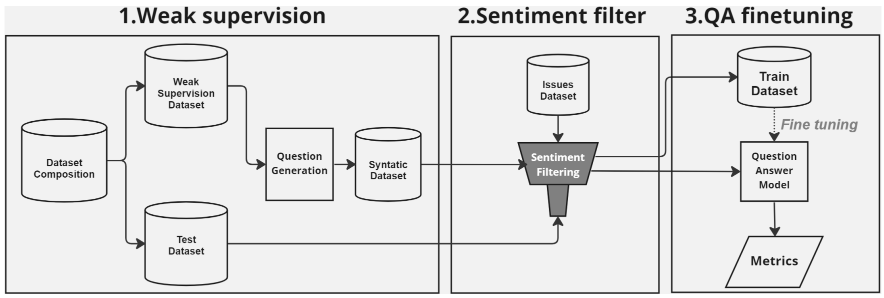

Fig. 1. Overview of the WSQASA method. The left side shows the weak supervision process, composed of a set of *datasets*, from which we split to establish a synthetic database generated by a QG model. At the center is a filtering process of questions and answers, carried out through a sentiment model and a correlation between the synthetic answers and a sentiment lexicon. Finally, the last block defines the process of tuning the QA model, comparing it to the same model without finetuning with WSQASA on the test database.

ically [11]. In QG, a natural language text is taken as input, and keywords are extracted to create factual questions.

The QG problem can be formulated as follows: given a sentence X containing n words and a target response A within X, the objective is to generate a question Y that leads to the answer A when considering the sentence X. More formally, the goal is to find the best question $\hat{Y}$ that maximizes the conditional likelihood given the passage X and the answer A [21]:

$$\hat{Y} = \arg\max_{Y} P(Y|X, A) \tag{2}$$

We utilize the training split's context to build the synthetic dataset and employ QG. The T5 model is used as a generative model to generate questions and answers. This model is trained on the SQuAD dataset, and its text-to-text approach allows for the end-to-end generation of questions. This process is similar to the GPT model [16]. In more detail, we employ T5 using the *text-to-text* approach, where the initial input is separated into sentences, and those with an associated response are highlighted using special *tags*. A token is added for each sentence with a response to indicate the extracted response. Figure 2 illustrates these steps for question extraction, where $< hl >$ and $< sep >$ tags indicate sentence separation and answers, respectively.

Finally, it is necessary to filter the generated questions and answers to ensure that only the ones containing information associated with sentiment analysis are used for the subsequent steps.

3.2 Sentiment Filter

The sentiment filter can be considered a "label function" that classifies the triples of questions, context, and answers based on their associated sentiment. To this

1. Python is a programming language. Created by Guido van Rossum and first released in 1991.
2. Python is a programming language. <hl> **Created by Guido van Rossum and first released in 1991** <hl>.
3. Python is a programming language. <hl> Created by Guido van Rossum and first released in <sep>**1991**<sep> <hl>.

Fig. 2. We follow a three-step process. The initial context is fed into T5 as the input in the first step. Then, the sentences containing an associated answer are separated, indicated by using the $< hl >$ tag. Finally, the response associated with each sentence is extracted using the $< sep >$ tag

end, we employed the Lexical Issue Dataset (LID)[1], which comprises well-formed sentences expressing negative sentiment polarity. The LID was constructed by applying lexical rules to datasets from online sources such as Reddit, Amazon, and news articles [10].

The filter we apply using *LID* can be formalized as follows: given a response whose vector representation is $a = a_0, a_1, \cdots, a_n$, and an embedding of an element from the *LID*, $b = b_0, b_1, \cdots, b_n$, we compute the cosine similarity using the Equation [23],

$$similarity(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}, \tag{3}$$

in which, $(\cdot)$, denotes the dot product of the two vectors, $|\cdot|$ denotes the norm, and n is the number of elements in each vector. We define a ϵ similarity threshold for QG data to be considered a new labeled data.

Table 3.2 presents a sample of the data filtering process. The first column represents the answers generated by the QG model, while the second column shows the responses filtered by similarity with the LID. The third column represents the sentiment filter associated with the LID, where the cut words are the ones removed by the filter. The table shows that the filter with the LID removes words such as dates, numbers, names, and other sentences that are not relevant to the domain.

Answer	**LID**	**LID + Sentiment Filter**
Department of Arizona	~~Department of Arizona~~	~~Department of Arizona~~
July 11, 2015	~~July 11, 2015~~	~~July 11, 2015~~
Afghanistan	Afghanistan	~~Afghanistan~~
Syria	Syria	~~Syria~~
hard beginning	hard beginning	hard beginning
difficulty reading and writing	difficulty reading and writing	difficulty reading and writing

Table 2. Representation of the filtering process over the dataset.

[1] https://github.com/rmarcacini/LID

The generation of questions and the subsequent sentiment filter results in a synthetic dataset for weak supervision. In the following section, we provide further details on how to perform fine-tuning of a QA model for sentiment analysis.

3.3 Question and Answering Fine-tuning

Following the steps above, we have generated a synthetic QA database focusing on sentiment analysis. Mishra et al. [19] have demonstrated that DistilBERT is a promising model for sentiment analysis through QA. Specifically, for sentiment analysis on tweets, DistilBERT outperforms more sophisticated models such as RoBERTa and AlBERT [15]. Our WSQASA method employs a QA version of DistilBERT for fine-tuning sentiment analysis.

The process of fine-tuning a QA-based sentiment model is described below. DistilBERT consists of six transformer encoders that can be fine-tuned for a specific task. With regards to the network loss function, given a sentence composed of n terms $S = s_1, s_2, \cdots, s_n$, the model generates an embedding associated with each term of the sentence $E = e_1, e_2, \cdots, e_n$. Each embedding $e_i = e_{i1}, \cdots, e_{im}$ is multiplied by a vector of initial weights $W = w_1, \cdots, w_m$, where $m = 512$ is the size of the embedding generated by DistilBERT. The dot product between each embedding and the weight vectors at the beginning and end of the sentence is then calculated, followed by a softmax applied to the result. Finally, the loss for the response start position is calculated as follows:

$$\begin{aligned} x &= \text{Concat}(e_1 \cdot W, \cdots, e_i \cdot W, \cdots, e_n \cdot W) \\ \hat{y} &= \text{softmax}(\mathbf{x_i}) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \\ H(y, \hat{y}) &= -\sum_{i=1}^{n} y_i \log \hat{y}_i \end{aligned} \tag{4}$$

where $Y = \{y_1, \cdots, y_n\}$ is the label associated with the correct positions of the *start* token, which the generative model gives. For the *end* weight vector, we need to change the W vector to the end weight*embedding* in the equation. The last equation is the loss function that we use to adjust the neural network weights.

In brief, the process of fine-tuning a DistilBERT model for QA-based sentiment analysis involves generating a synthetic dataset through the combination of question generation and data filtering. The model is then fine-tuned by modifying the weights of its transformer encoders using a loss function that calculates the probability of the correct start and end positions for the sentiment answer in a given text. DistilBERT has been shown to outperform more complex models like RoBERTa and AlBERT for sentiment analysis on tweets, making it a promising choice for this task. Overall, the fine-tuned QA-based sentiment analysis model can be a useful tool for analyzing sentiment in various applications, as discussed in the next section.

4 Experimental Evaluation

This section presents two specific experiments to provide a general understanding of the method. We employ T5 for question generation for each of these experiments, followed by a domain-specific filter and fine-tuning of the DistilBERT QA model. The QA model is trained for a single epoch with a learning rate of $2^{10^{-5}}$ and weight decay of 0.01.

We utilize various datasets, including TweetQA [28] consisting of tweets extracted from news, CoQA [24] comprising conversations written in natural language, Adversarial QA [2] created through the adversary use of models in the annotation process, QED [13] designed to enhance the interpretability of questions and answers, DuoRC [25] generated by extracting questions and answers under different contexts that comment on the same subject, and DROP [7] that seeks to establish models capable of discrete operations. However, this work disregards these datasets' unanswered answers and non-extractive questions. Specifically, we use six datasets of extractive QA for cross-domain training. We test the QA model on one of these datasets while using the contexts of the others as input to the generative model.

One of the relevant points regarding the proposed *framework* refers to the choice of model for the sentiment filter. In this sense, as discussed throughout Section 3.2, models based on *transformers* tend to take contexts into account, which is reflected in a broader coverage by the model.

One relevant aspect of the proposed method concerns the use of a sentiment filter combined with the QG model. Figure 3 illustrates the results obtained in this scenario. It is worth noting that the similarity filter, in conjunction with the lexical filter, improved the model's understanding of the task, resulting in significantly better results for all the analyzed datasets. The LID similarity filter was able to eliminate malformed questions and answers generated by *T5*, thus reducing the noise in the synthetic training dataset, while the lexical sentiment filter made the sentiment analysis task more explicit.

Figure 3 presents the ***relativegain*** that the DistilBERT model achieved when applying WSQASA on various *datasets* represented along the x axis. The dark gray bar indicates the *accuracy*, while the light gray represents the *f1-score*. The relative gain for $f1_{gain}$ and acc_{gain}, where $f1_{wsqasa}$ and acc_{wsqasa} are the $f1-score$ and accuracy of DistilBERT after WSQASA tuning, and $f1_{distil}$ and acc_{distil} are the DistilBERT performance without WSQASA tuning, can be calculated using the following equations:

$$
\begin{aligned}
f1_{gain} &= f1_{wsqasa} - f1_{distil} \\
acc_{gain} &= acc_{distil} - acc_{distil}
\end{aligned}
\tag{5}
$$

If $f1_{gain}$ is positive, it indicates that the WSQASA has improved the performance of the QA model concerning the $f1$-score measure. Similarly, if acc_{gain} is positive, the WSQASA has improved the performance of the QA model concerning the accuracy measure.

Our method shows an overall improvement in the performance of the model for most of the datasets analyzed. However, in some cases, there was a slight decrease in the model's performance in terms of accuracy and $F1$-score, respectively. This may be attributed to the challenges posed by these datasets, as well as the lack of a correlation filter with the LID dataset.

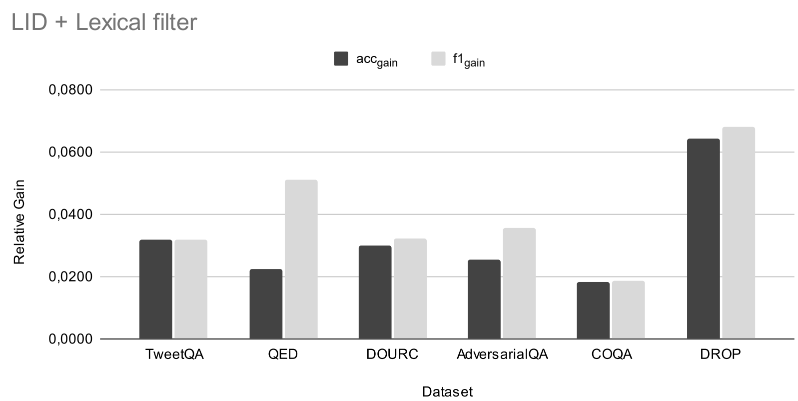

Fig. 3. Results after applying the WSQASA method to the DistilBERT model, using the correlation filter with the LID and lexical sentiment filter.

Figure 4 presents a general comparison of the proposed method. The bubble plot illustrates the size of the test dataset, with the radius of each bubble representing its size and the color representing the different experiments conducted. The test datasets under analysis are displayed along the x axis, while the red line represents the baseline for this analysis, which corresponds to the performance of DistilBERT without the application of WSQASA. All experiments that show a bubble above this line are considered to improve the QA model.

We standardized the names of the experiments based on the filters applied in the second part of the WSQASA pipeline. The first filter, $none + none + RoBERTa$, was selected as the baseline and used *RoBERTa* without any correlation with the LID. The other filters include the ε value associated with the correlation with the LID, followed by the source from which the data were extracted for the construction of the LID and the type of sentiment filter applied to the dataset. For example, the experiment name *LID:0.7+amazon+Lexical* evaluates the similarity filter with the *amazon* source of the LID, which has a similarity of at least $\varepsilon = 0.7$, followed by the application of a lexical sentiment filter. In cases where the source is *all*, such as in *LID:0.7+all+Lexical*, all LID sources were used.

Our results indicate that the proposed method significantly improves the QA model in almost all cases, as most bubbles lie above the baseline. The optimal

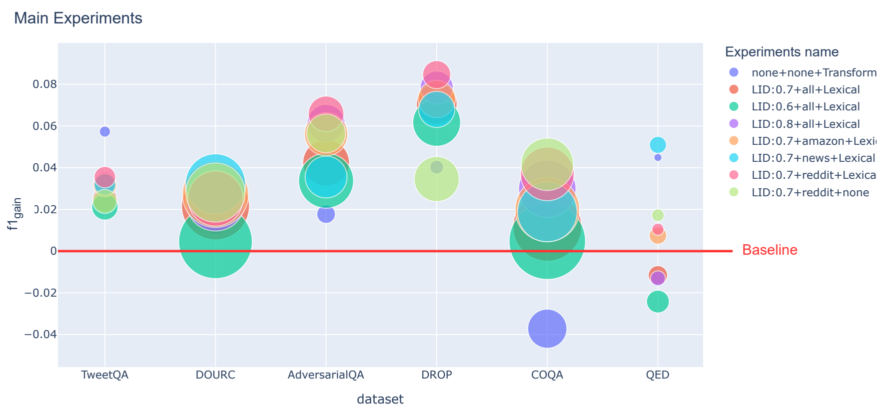

Fig. 4. Comparison chart of the different approaches to the WSQASA, the radius of the bubbles represents the size of the test dataset, while the axis y presents the relative gain of *F1*, the axis x represents the datasets under analysis and the colors of the bubbles represent the experiments carried out.

ε for the proposed method should be between 0.7 and 0.8 to ensure that the model has sufficient data for fine-tuning. The results above demonstrate that the proposed WSQASA method can substantially enhance QA models for particular domains, even when labeled data is unavailable. This section has shown the method's versatility and how it can be tailored to yield improved outcomes for QA models.

5 Conclusion

In this paper, we introduced a novel method, WSQASA, aimed at addressing the lack of labeled data in QA-based sentiment analysis tasks. Our approach leverages weak supervision through QG and domain-specific filtering to improve QA models, even in scenarios without labeled data. We demonstrated the effectiveness of our approach on six different datasets, reporting significant improvements in model performance.

The flexibility of WSQASA allows for customization and improvement of different modules. Future research could explore other transformer-based QA models, such as DeBERTa-v3 [9], and investigate the application of other QG models, such as GPT [4]. While this paper focuses on sentiment analysis, WSQASA has the potential to be applied to other domains and tasks.

In conclusion, our proposed method provides a promising solution to the challenges posed by the lack of labeled data in QA-based sentiment analysis tasks. By leveraging weak supervision and domain-specific filtering, WSQASA significantly improves model performance. We hope this work will inspire further research in this area and lead to the development of more effective and efficient approaches for sentiment analysis and other NLP tasks.

References

1. Banerjee, P., Gokhale, T., Yang, Y., Baral, C.: Weaqa: Weak supervision via captions for visual question answering. arXiv preprint arXiv:2012.02356 (2020)
2. Bartolo, M., Roberts, A., Welbl, J., Riedel, S., Stenetorp, P.: Beat the ai: Investigating adversarial human annotation for reading comprehension. Transactions of the Association for Computational Linguistics **8**, 662–678 (2020)
3. Bengio, Y., et al.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1), 1–127 (2009)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering **27**(4), 1071–1092 (2020)
6. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. arXiv preprint arXiv:1708.06022 (2017)
7. Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M.: Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. arXiv preprint arXiv:1903.00161 (2019)
8. Hao, T., Li, X., He, Y., Wang, F.L., Qu, Y.: Recent progress in leveraging deep learning methods for question answering. Neural Computing and Applications pp. 1–19 (2022)
9. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
10. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517 (2016)
11. Indurthi, S.R., Raghu, D., Khapra, M.M., Joshi, S.: Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 376–385 (2017)
12. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology **60**(11), 2169–2188 (2009)
13. Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E., Soares, L.B., Collins, M.: Qed: A framework and dataset for explanations in question answering. Transactions of the Association for Computational Linguistics **9**, 790–806 (2021)
14. Lee, K., Salant, S., Kwiatkowski, T., Parikh, A., Das, D., Berant, J.: Learning recurrent span representations for extractive question answering. arXiv preprint arXiv:1611.01436 (2016)
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

16. Lopez, L.E., Cruz, D.K., Cruz, J.C.B., Cheng, C.: Transformer-based end-to-end question generation. arXiv preprint arXiv:2005.01107 **4** (2020)
17. Mei, Q., Radev, D.R.: Information retrieval (2014)
18. Melville, P., Gryc, W., Lawrence, R.D.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1275–1284 (2009)
19. Mishra, C., Bansal, E., Victoria, A.H.: Tweet sentiment extraction. Tech. rep., EasyChair (2022)
20. Nassiri, K., Akhloufi, M.: Transformer models used for text-based question answering systems. Applied Intelligence pp. 1–34 (2022)
21. Pan, L., Lei, W., Chua, T.S., Kan, M.Y.: Recent advances in neural question generation. arXiv preprint arXiv:1905.08949 (2019)
22. Perrin, A.: Social media usage. Pew research center **125**, 52–68 (2015)
23. Rahutomo, F., Kitasuka, T., Aritsugi, M.: Semantic cosine similarity. In: The 7th international student conference on advanced science and technology ICAST. vol. 4, p. 1 (2012)
24. Reddy, S., Chen, D., Manning, C.D.: Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics **7**, 249–266 (2019)
25. Saha, A., Aralikatte, R., Khapra, M.M., Sankaranarayanan, K.: Duorc: Towards complex language understanding with paraphrased reading comprehension. arXiv preprint arXiv:1804.07927 (2018)
26. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
27. Silva, E.H.d., Marcacini, R.M.: Aspect-based sentiment analysis using bert with disentangled attention. In: Proceedings (2021)
28. Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Chang, S., Guo, X., Wang, W.Y.: Tweetqa: A social media focused question answering dataset. arXiv preprint arXiv:1907.06292 (2019)
29. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1253 (2018)
30. Zhou, Z.H.: A brief introduction to weakly supervised learning. National science review **5**(1), 44–53 (2018)