machine learning & knowledge extraction

MDPI

Article

The Case of Aspect in Sentiment Analysis: Seeking Attention or Co-Dependency?

Anastazia Žunić , Padraig Corcoran and Irena Spasić *

School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, UK; zunica@cardiff.ac.uk (A.Ž.); corcoranp@cardiff.ac.uk (P.C.)
* Correspodence: spasici@cardiff.ac.uk

Abstract: (1) Background: Aspect-based sentiment analysis (SA) is a natural language processing task, the aim of which is to classify the sentiment associated with a specific aspect of a written text. The performance of SA methods applied to texts related to health and well-being lags behind that of other domains. (2) Methods: In this study, we present an approach to aspect-based SA of drug reviews. Specifically, we analysed signs and symptoms, which were extracted automatically using the Unified Medical Language System. This information was then passed onto the BERT language model, which was extended by two layers to fine-tune the model for aspect-based SA. The interpretability of the model was analysed using an axiomatic attribution method. We performed a correlation analysis between the attribution scores and syntactic dependencies. (3) Results: Our fine-tuned model achieved accuracy of approximately 95% on a well-balanced test set. It outperformed our previous approach, which used syntactic information to guide the operation of a neural network and achieved an accuracy of approximately 82%. (4) Conclusions: We demonstrated that a BERT-based model of SA overcomes the negative bias associated with health-related aspects and closes the performance gap against the state-of-the-art in other domains.

Keywords: sentiment analysis; natural language processing; deep learning; transformers; syntactic dependencies

check for updates

Citation: Žunić, A.; Corcoran, P.; Spasić, I. The Case of Aspect in Sentiment Analysis: Seeking Attention or Co-Dependency? *Mach. Learn. Knowl. Extr.* **2022**, *4*, 474–487. https://doi.org/10.3390/make4020021

Academic Editor: Ján Paralič

Received: 15 March 2022
Accepted: 9 May 2022
Published: 13 May 2022

1. Introduction

Sentiment analysis (SA), also known as opinion mining, is an area of natural language processing (NLP) that focuses on the classification of the sentiment that is expressed in a written document. Formally, SA is defined as the task of identifying a quadruple (s, g, h, t) whose values represent the sentiment, the object targeted by the sentiment, the holder of the sentiment and the time at which the sentiment was expressed [1]. In practice, SA has traditionally focused on a simpler task of finding the pair (s, g). Here, the sentiment s is often conflated to polarity [2], which can be either positive or negative, even though other classification schemes can be utilised [3]. The target g has typically been the overall topic of an analysed text document. In principle, the target is some entity, but can also be an aspect of such an entity. Here, an aspect represents some characteristic of such an entity [4]. The choice of these characteristics depends on a specific domain in which SA is applied. Aspect-based SA refines the focus of SA by classifying the sentiment associated with a specific aspect and not just the overall sentiment associated with the entity.

By lowering the barrier to entry, Web 2.0 solicited user feedback on institutional platforms and encouraged user interaction on social media, giving rise to vast amounts of user-generated content. This in turn presented an opportunity to study and monitor public opinion in real-time by applying techniques from NLP, which led to SA becoming one of the fastest growing areas of research in this space. Typically, text sources used to facilitate such research originate from social media or customer reviews. Web 2.0 also led to a proliferation of health-related online platforms, even more so in the most recent times, during which access to healthcare has been restricted for a wide range of possibly undiagnosed medical conditions. In particular, many users resorted to self-medication, exchanging information about pharmaceutical drugs online. Not surprisingly, much of the research into SA in relation to health and well-being focuses on drug reviews. Aspect-based SA of such reviews can in turn be used to support pharmacovigilance by detecting adverse drug reactions [5]. The most obvious aspects in this case would be drug indications and side effects. For instance, consider the following examples in which the word headache represents an aspect:

1. It's the only drug that works for my $\underbrace{\text{headache}}_{\text{positive}}$.
2. A dose of 750 mg twice daily had no effect on my $\underbrace{\text{headache}}_{\text{negative}}$.
3. Caused vomiting and gave me the worst $\underbrace{\text{headache}}_{\text{negative}}$.
4. I find using a half a capsule seems to work fine without giving me a $\underbrace{\text{headache}}_{\text{positive}}$.

These examples illustrate how the sentiment towards the same aspect can vary across different contexts. Unlike their counterparts in other domains, e.g., quality and price in product reviews, aspects such as headache are a priori negative. Many SA tools struggle to counteract such bias and actually perform better when an otherwise negative aspect is removed from consideration [6]. In general, SA applications in health and well-being lag behind the state-of-the-art in other domains [7]. This can be attributed to the negative sentiment that gets associated with medical conditions by default, which requires careful analysis to disentangle such sentiment from that of its context. This cannot be achieved by conflating the context into a flat bag-of-words representation. Instead, a more sophisticated representation is required, together with an algorithm that can effectively process such a representation. Neural networks cannot only learn complex relationships between individual words in a sentence but can also utilise more complex sentence representations, such as graphs, which can be based on syntactic parses. For example, convolution can be successively applied to an aspect of SA by traversing through its syntactic dependencies [8]. More recently, language models that are pre-trained using transformer architectures [9] have demonstrated significant improvements over recurrent neural networks in a variety of NLP tasks, including natural language understanding, named entity recognition and question answering [10]. The most popular architecture of this kind is called Bidirectional Encoder Representations from Transformers (BERT) [10]. Its popularity lies in the fact that it cannot only be pre-trained to generate contextualised word embeddings, but can also be easily fine-tuned using relatively small datasets to support downstream NLP tasks such as that of SA.

In this study, we investigate the potential of a BERT-based approach to aspect-based SA in the domain of health and well-being. The remainder of this article is organised as follows. Section 2 provides an overview of related work. Section 3 describes the methodology, including data collection, implementation details and model training. In Section 4, we evaluate the model and compare it to the baseline established in a previous study. Section 5 discusses a possible interpretation of the results by analysing the internal operation of the model. Finally, Section 6 concludes the paper.

2. Related Work

Until recently, the vast majority of research in SA in health and well-being used rule-based and traditional machine learning techniques [7]. Both approaches employ simple features such as n-grams, which fail to capture relationships that are more complex than simple co-occurrence. Not surprisingly, deep learning models, which can learn to capture the semantics of individual words and complex relationships among them, tend to outperform traditional machine learning methods.

Various deep learning architectures have been built to support aspect-based SA. Most of them represent some variation of recurrent neural networks (RNNs), such as long short-term memory (LSTM), e.g., [11–13]. When it comes to aspect-based SA of drug reviews, a bidirectional gated recurrent unit (GRU) with an attention layer was proposed in [14]. RNNs are optimised to process sequences and, therefore, are not ideally suited for context-sensitive tasks such as aspect-based SA. Convolutional neural networks (CNNs) are better suited to represent contextual information and, as such, have been used in SA applications as local feature extractors [15,16]. Alternatively, an attention mechanism can be used to improve the performance of RNNs in aspect-based SA by letting them know where to focus their learning. An attention-based bidirectional CNN-RNN provides a hybrid model in which bidirectional RNNs are used to model both long and short contextual dependencies, local features robust to positional changes are selected using CNNs, and an appropriate emphasis is placed on different words by applying the attention mechanism on the output of bidirectional layers [17].

Aspect-based SA is a fine-grained task that aims to classify the sentiment towards a particular aspect. The aspect's relations to other words represent important features of its sentiment, but are not taken into account in RNN-based approaches. In our previous work, a graph convolutional network (GCN) designed to operate on syntactic dependencies outperformed the traditional RNN approach by a large margin on the task of aspect-based SA of drug reviews [8]. A GCN approach may not be able to capture the features of long-distance dependence, thus struggling to effectively represent the aspect's context. This issue can be easily resolved by adding transitive edges to the dependency graph, which has been proven to improve the representation of sentiment dependencies [18]. Alternatively, a phrase dependency graph can be constructed by integrating the constituency and dependency parse trees [19]. Further embellishing the dependency graph by leveraging information from a sentiment lexicon was found to improve the learning ability of a GCN model in aspect-based SA [20]. However, adding more information may introduce noise and inefficient use of information relevant to SA. Namely, despite the direct or indirect connection with an aspect in the dependency tree, only few words add value to predicting the sentiment polarity of the aspects. These words tend to be adjectives and verbs. Therefore, part-of-speech information can be used to prune the dependency tree, with two benefits [21]: First, fewer unrelated words are connected directly or indirectly to the aspect, which reduces the noise they bring to the convolution. Second, a more concise syntactic dependency graph leads to fewer convolutions, thus making the corresponding GCN more efficient. Apart from local dependencies, context is also a function of historical utterances. To place an aspect into a wider context, which may still be relevant for its sentiment, lessons can be learnt from conversational SA. The emotional recurrent unit provides a compact RNN architecture that encodes the context information, captures the influence of context information for a sentence and extracts features for sentiment classification [22].

SA suffers from domain dependency [2]. On the one hand, it requires a lot of training data. In particular, deep learning algorithms are known to be data hungry. On the other hand, when an SA model trained on one domain is applied to a different one without any transfer of knowledge, the performance tends to deteriorate. One way to tackle the problem of domain shift is to create an ensemble of models trained on different data sources [23]. An ensemble of models can combine individual predictions in a way that the given models compensate for each other's weaknesses [24]. In particular, heterogeneous ensembles use different learning algorithms to generate different types of base classifiers. Recent experiments in SA demonstrated that ensemble learning can improve the accuracy. For example, the stacking of LSTM, CNN and CNN-BiLSTM and support vector machine (SVM) significantly improved the accuracy of SA in Chinese, albeit failing to replicate the success in English [25]. Nonetheless, another study, which widened the choice of base classifiers to four pre-trained, lexicon-based models and six machine learning algorithms (naïve Bayes, SVM, logistic regression, feedforward neural network, CNN and LSTM) managed to improve performance by more than five percentage points over the best individual model [24]. The true potential of ensemble approaches to SA lies in leveraging symbolic models (such as lexicons and grammatical relationships) to encode meaning and subsymbolic methods (such as word embeddings and neural networks) to infer patterns from data [2].

More recently, research attention has shifted towards large, pre-trained language models such as BERT [10]. BERT is a transformer-based architecture that provides contextual word embeddings; it uses an attention-based mechanism, rather than recurrence, to determine which words are important for the overall context within the document [9]. It has enabled great performance improvements across a variety of NLP tasks. The main advantage of BERT is that it can easily be fine-tuned using additional training data to solve specific NLP tasks, such as aspect-based SA. This task can be formulated as a question-answering task, where the aspect represents a question and its sentiment is the answer. BERT typically represents this task by pairing up two sequences, one representing the source sentence and the other one specifying the phrase that corresponds to the aspect. This approach was successfully adapted to classify the sentiment associated with a specific aspect of a product or a restaurant as expressed in user reviews [26–29]. Such an approach improved the results relative to models that use a single sequence to perform aspect-based SA [26].

BERT is commonly pre-trained for specific domains to improve its performance on different sublanguages. For example, of relevance to the domain of health and well-being are BioBERT [30] and ClinicalBERT [31]. However, when lay language is processed in this domain, BERT's performance may still be superior to the specially trained language models. For example, when BERT was used to understand people's opinion towards vaccination, a multilingual BERT model outperformed both BioBERT and ClinicalBERT [32]. BERT was successfully fine-tuned to perform SA of drug reviews [33], albeit without focusing on specific aspects. In this study, we investigate the application of BERT to aspect-based SA of drug reviews. We also relate the results to the ones achieved in our previous study, which was based on graph convolution over the dependency graph [8].

3. Methodology

The goal of aspect-based SA is to classify the sentiment of a document with respect to a particular aspect. Therefore, the document and the aspect considered constitute the input, whereas the output represents the sentiment, classified into one of two classes, positive or negative. Here, neutral sentiment is ignored in line with the vast majority of SA approaches, which ignore neutrality despite the evidence that it is the key for distinguishing between the two polarities [34,35]. However, the reason for ignoring neutrality in this study is not pragmatic. Instead, this decision was related to the choice of aspects in this study. Namely, we focused on signs and symptoms as the aspects of SA.

A sign is an objective observation of a potential health issue. A symptom is a subjective experience of a potential health issue. Given that both signs and symptoms indicate a potential health issue, they bear a negative sentiment. When taking a pharmaceutical drug to address a given health issue, there are three possibilities for the associated signs and symptoms. They can remain unchanged, they can worsen, or they can improve. In the first two cases, the sentiment associated with a sign or symptom remains negative. In the latter case, one may assume that the resolution of a sign or symptom leads to a better state of health and thus turns the underlying negative sentiment into a positive one. In other words, when it comes to health, no one is expected to be ambivalent; hence, we decided not to consider neutral sentiment.

Before analysing their sentiment, all aspects are identified automatically by matching concepts classified as signs or symptoms in the Unified Medical Language System (UMLS), a large repository of inter-related biomedical concepts and the corresponding terminology [36].

3.1. Neural Network Architecture

The first step in fine-tuning BERT for aspect-based SA is choosing an appropriate representation of the problem. BERT [10] is a transformer-based language model. Unlike RNNs, which iterate through sequences, transformers are based on an encoder–decoder NN architecture that uses an attention mechanism to support the holistic interpretation of a sequence [9]. The self-attention layer considers all words, each represented by its embedding and its position relative to other words, to improve its encoding of the entire sentence. In other words, self-attention determines the impact of individual words on sentence interpretation. During training, BERT hides a certain percentage of the words by using a special token (MASK) instead and uses their position to infer these words. This serves to prevent the decoder from looking ahead when predicting the next word, thus effectively making the training parallel. By performing this task, BERT learns the relationships between words.

Masked language modelling is only one of two tasks on which BERT is trained simultaneously. The second task is next sentence prediction, which allows BERT to learn long-term dependencies across sentences. BERT uses two special tokens to support fine-tuning and specific task training. The first one is a classification token (CLS). It indicates the beginning of a segment, typically a sentence, and is commonly used for classification tasks, hence the name. The output associated with this token is used to make a prediction about the given segment. The other special token is a delimiter token (SEP). It simply indicates the end of a segment. The type of segments used depends on the specific task BERT is fine-tuned for. For instance, in question-answering, one segment can be a question, whereas the other one can be the reference text. The two segments are then appended and separated by a special delimiter token (SEP).

In our model, we chose the aspect of SA as one segment and its context (i.e., the whole sentence) as the other. This can be seen at the top of Figure 1, which illustrates the architecture of a BERT-based model for aspect-based SA. In this example, the sentence and the aspect in question are "This medicine works great when it comes to pain" and "pain", respectively, which are combined into the following input sequence: "(CLS) This medicine works great when it comes to pain (SEP) pain (SEP)" (see the yellow row in Figure 1). Finally, to meet the fixed-length requirement BERT expects of its input, such a sequence is padded using a special token (PAD) until the maximum length of 70 tokens has been reached.

The input sequence is then processed as follows. First, each token's vocabulary identifier is mapped to a token embedding that was learned during training (see the orange row in Figure 1). A binary vector is then used to differentiate between two text segments. The binary vector is mapped to a segment embedding (see the grey row in Figure 1) using a lookup table, which was also learned during training. Finally, local token positions are mapped to positional embeddings (see the blue row in Figure 1) using a lookup table, which was updated during training.

Similarly to binary classification tasks described originally in [10], the final transformer output that corresponds to the special (CLS) token amounts to an aggregate problem representation, i.e., a pooled output. To determine the sentiment from this aggregate representation of a sentence and its aspect, the pooled output is fed into the classification layer (see the pink rows in Figure 1). The classification layer reduces the size of the pooled output to two dimensions, which correspond to the log-odds (or logits) of the classification output with respect to the question of whether the implied sentiment is positive or negative. The classification layer is not pre-trained, unlike the preceding layers of the NN. Multiple pre-trained BERT models can be used here. They differ with respect to the choice of hyperparameter values. We employed the $\text{BERT}_{\text{base}}$ model, which was pre-trained using 12 layers of transformer encoders, 12 attention heads and the hidden dimension of 768. Going back to the classification layer, its output is passed through the softmax function, which estimates the probability distribution over positive and negative sentiments.

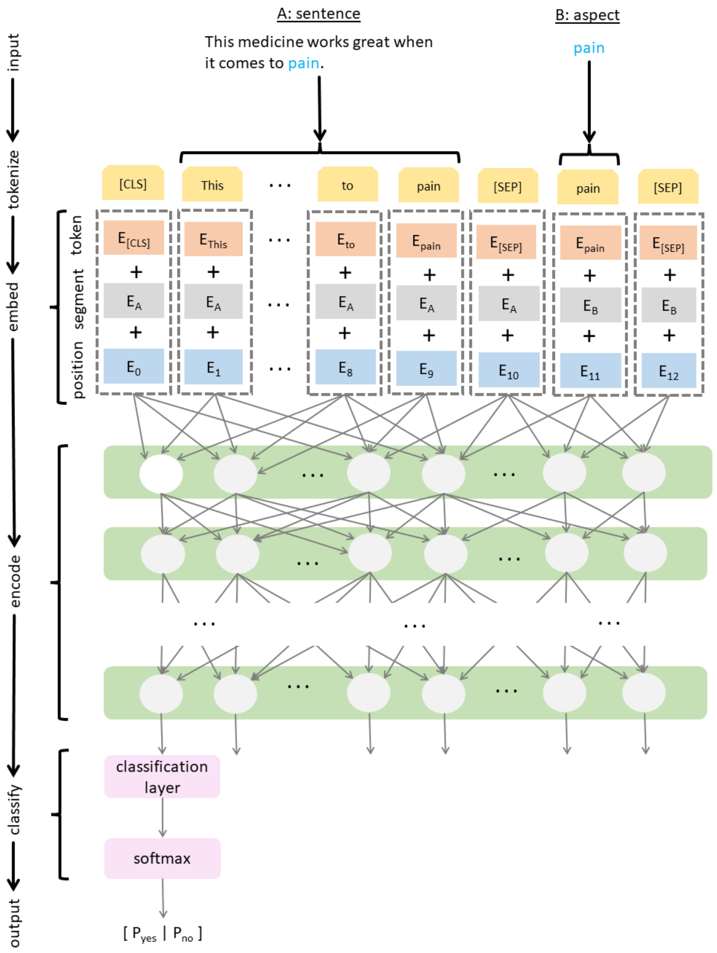

Figure 1. BERT-based architecture for aspect-based SA.

3.2. Implementation and Training

To implement our approach described above, we used the publicly available pre-trained BERT$_{\text{base}}$ model. Specifically, we used its distribution from Hugging Face, an open-source library that consists of state-of-the-art transformer architectures under a unified API [37]. The pre-trained BERT model was fine-tuned by minimising cross-entropy loss, which is calculated between the output from the softmax and the true labels. The loss function was optimised with Adam optimizer [38], a stochastic gradient-descent method that is based on adaptive estimation of first-order and second-order moments, with the learning rate set to 2×10^{-5}. The specific learning rate was selected based on the suggestions made in [10]. All other hyperparameters were set to their default values.

The classification model was trained for four epochs following the recommendations of BERT's original authors to use 2–4 epochs to fine-tune BERT for a specific NLP downstream task [10]. We evaluated the model on the validation set after each epoch. During each epoch, the model parameters were updated with respect to the error of each batch of the training data. Batch size for training, validation and test sets was set to 16.

The overall SA system was implemented in Python programming language using PyTorch [39], a deep learning framework that combines usability and speed by coding executable models, thus making debugging easier while being efficient and supporting further hardware acceleration. All our experiments were run on the CPU, not the GPU, of a PC with an Intel processor with 6 cores, each running at 2.6 GHz, and 16 GB RAM.

4. Results

4.1. Data

To train the model and evaluate its performance, we used a dataset we created specifically for the task of aspect-based SA. It was described in detail in a previous study, which proposed deep learning over the syntactic dependency graph using graph convolution [8]. Here we summarise its basic properties. It consists of drug reviews borrowed from another study [40], which were publicly available from the UCI Machine Learning Repository. These reviews were originally collected from the Drugs.com website [41]. Each review comes with a star rating on a scale from 1 to 10, which was converted into a sentiment label.

To adapt this dataset for the task of evaluating aspect-based SA, aspects were automatically annotated by matching concepts classified as signs or symptoms in the UMLS [36]. To increase the likelihood of the overall sentiment being related to a specific aspect considered, only short reviews, specifically those consisting of a single sentence, were considered. The final dataset is comprised of 1232 sentences, out of which 639 and 593 sentences were labelled with positive and negative sentiment, respectively. Further details on this dataset are available in our previous study [8], whose results were also used to establish the baseline for this study.

The dataset was split randomly so that 80% and 20% of the data were used for training and testing respectively. Further, approximately 20% of the training data were used to validate the model trained on the remainder of the training data. Table 1 illustrates the distribution of data across the three subsets (training, validation and test set) and the two sentiment labels (positive and negative). No substantial differences in the distribution of the two labels across the three datasets can be observed.

Table 1. The distribution of sentiment labels across the three datasets.

	Positive	Negative	Total
Train	410	378	788
Validation	99	98	197
Test	130	117	247
Total	639	593	1232

4.2. Evaluation

To be able to make the direct comparison to the baseline results, we reused the evaluation measures described in the previous study [8]. Specifically, we used measures commonly used to evaluate classification performance, including accuracy and cross entropy loss. Accuracy represents the percentage of correctly classified instances. Accuracy is not always a reliable metric. For example, it may provide misleading results when the test dataset is unbalanced. As we can see from Table 1, this is not the case in this study, thus justifying the use of this metric. Given that our model also provides probability distribution over the sentiment labels as output, we used cross entropy loss to compare the predicted probabilities to the gold standard labels as follows:

$$loss = -\frac{1}{n}\sum_{i=1}^{n} ln(p_i) \qquad (1)$$

where p_i is the corrected probability, i.e., the probability that a particular prediction matches the gold standard label. The closer the predicted probability to the gold standard label, the lower the cross entropy loss.

In addition to running experiments using the standard BERT model, we performed experiments with its distilled version. DistilBERT is a smaller, general-purpose language model, which can be fine-tuned for specific tasks just like its larger counterpart [42]. It reduces the size of a BERT model by 40% while retaining 97% of its language understanding capabilities with the benefit of being 60% faster to run. Both language models come in both cased and uncased versions. In the uncased models, the text is lowercased prior to WordPiece tokenization, thus making the model case-insensitive. No case changes are performed on text in the cased version.

Table 2 provides the results. The significance of case can be immediately observed. In both BERT and DistilBERT, the cased model outperformed the uncased one by a large margin, with accuracy in the 70s and 90s, respectively. In fact, the uncased models performed worse than the baseline. On the other hand, the cased models achieved an accuracy of almost 95%, outperforming the baseline by more than 12 percentage points. The difference in performance between the two cased models was negligible.

Table 2. The evaluation results.

Method		Accuracy	Loss
Baseline		81.78%	0.4570
BERT$_{\text{base}}$	uncased	78.14%	0.5270
	cased	94.33%	0.3641
DistilBERT$_{\text{base}}$	uncased	73.28%	0.5688
	cased	94.74%	0.3660

5. Discussion

5.1. Error Analysis

The impact of casing on the performance of SA was unexpected, so it warrants further analysis to try to explain this phenomenon. Intuitively, one might expect this issue to be related to the use of the personal pronoun I, which is often used to describe one's state. In particular, within the realm of health and well-being, the usage of pronouns was found to have an effect on SA even more so than on standard English usage [43]. The total number of sentences that were correctly classified by the cased model but incorrectly classified by the uncased model was 47. Therefore, error analysis was not a laborious undertaking. Table 3 provides a sample of errors made by the uncased model that were corrected by the cased model. For simplicity, the results in this section are based on the standard BERT model.

Table 3. A sample of sentences incorrectly classified by the uncased model but correctly classified by the cased model.

ID	Sentence	Label	Uncased	Cased
1	Excellent **headache** reliever!	+	−	+
2	Good medicine, it gets rid of your **pain** without that drowsy sick feeling.	+	−	+
3	Love this medicine, no **headache**.	+	−	+
4	Sadly no effect on my **pain**.	−	+	−
5	Made my **symptom** worse-so much for 24 h relief.	−	+	−
6	No **pain** relief whatsoever.	−	+	−

Within 47 sentences, we found only 10 mentions of the personal pronoun 'I' that were not at the beginning of the sentence. In the majority of cases, such as those shown in Table 3, we can see that the personal pronoun 'I' did not play any role in these sentences, so we can dismiss our initial hypothesis and investigate other possible effects of casing on the classification performance. In the same table, the words highlighted using a bold typeset represent the aspect. The case of an aspect was clearly not affected by lowercasing. In fact, none of the other words were affected by lowercasing apart from the first word of a sentence. English grammar requires the first word of a sentence to be capitalised. A quick inspection of the first words reveals the majority to be emotionally charged words that are typically found in most sentiment lexica, e.g., excellent, good, love and sadly. We inspected all errors and indeed found that in all such cases the correct sentiment of the whole sentence coincided with the sentiment of these words. When the model was pre-trained, it was reasonable to assume that these words were also found at the beginning of a sentence, as there are few other cases that would require their capitalisation. Therefore, their learnt embeddings would be correlated with their initial position in a sentence. When the model is fine-tuned, these words also have the most immediate impact on the neighbouring special token (CLS), which represents a pooled output. Therefore, it is reasonable to assume that the performance of the model is more directly linked to the position of these words rather than their casing alone.

The baseline model was not case sensitive. It also used convolution relative to the aspect of a sentence. It was, therefore, less influenced by the initial word, unlike the BERT model that uses a pooled output that is associated with a special token positioned before the start of a sentence. Nonetheless, BERT outperformed the baseline approach.

5.2. Model Interpretability

To investigate the internal logic of the BERT model, we used Captum [44], an open-source library for model interpretability. It uses integrated gradients [45], an axiomatic attribution method that attributes the prediction of a deep neural network to its inputs. Two fundamental axioms that an attribution method should satisfy ensure that any artefacts affecting the attribution method are related to either the data or the neural network rather than the method itself. The first axiom, sensitivity, states that (1) whenever input and baseline differ in only one feature but have different predictions, then that particular feature should be given a non-zero attribution, and (2) if the function implemented by the neural network does not depend on some variable, then that particular variable should be always be given zero attribution. The second axiom, implementation invariance, states that any two functionally equivalent networks should receive identical attributions regardless of any differences in their implementations.

Of note, this attribution method only measures the relative importance of features in a neural network but does not address the interactions between the features nor the internal logic of the network. To study the extent to which syntactic dependencies between an aspect and other tokens (i.e., features in this context) are correlated with the attributions assigned to these tokens, we cross-referenced the attribution scores received by each token to their distance from the aspect in the syntactic dependency graph. Figure 2 provides an example of cross-referencing a token's distance from the aspect to its attribution score. The zero distance indicates the aspect, in this case the word 'pain'. The attribution score has been colour-coded using the heatmap colour palette given on the right. In this particular example, the tokens that are one to two steps away from the aspect in the dependency graph received the highest attribution score, with the exception of the punctuation token. Of note, the closest token to the aspect's right in the sequence graph (i.e., the word 'and') received the lowest score. A similar trend continues to the right, including the positive word 'help' also receiving a low attribution score and thus not contributing significantly to the positive sentiment of the aspect 'pain'.

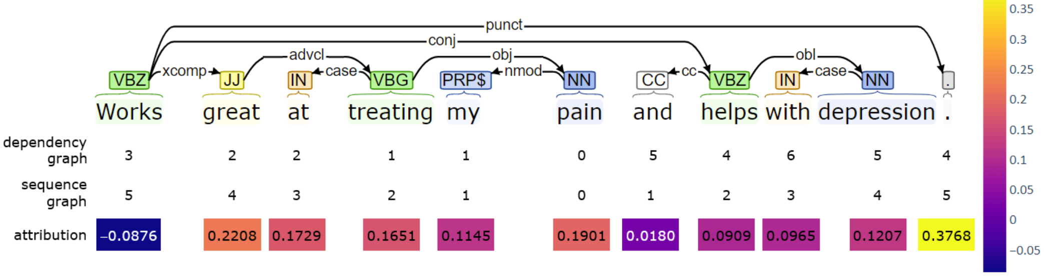

Figure 2. An example of cross-referencing a token's distance from the aspect in the dependency and sequence graphs, respectively, to its attribution score.

5.3. Statistical Analysis

To see whether this anecdotal evidence can be generalised, we performed statistical analysis to check whether higher attribution scores are correlated with smaller distances in the dependency graph. We used the Pearson correlation coefficient, which is calculated according to the following formula:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}, \tag{2}$$

where x_i represents the i-th data point in vector x, whereas $\overline{x}$ represents the mean value of vector x. Here, the null hypothesis is that there is no correlation between the attribution score and the distance of the token from the aspect in the dependency graph. The correlation between the two variables was found to be -0.074. In other words, the smaller the distance, the higher the attribution score and vice versa. The corresponding p-value of 5.4732×10^{-19} was smaller than the set threshold of 0.05, indicating that the correlation between the two variables was statistically significant. Therefore, the null hypothesis was rejected.

We performed analogous experiments using the sequence graph representation, i.e., we measured the correlation between the token attribution score and the distance of the token from the aspect in the sequence graph. The correlation between these two variables was found to be -0.069 with a p-value of 1.4107×10^{-16}. It came as no surprise that the local context of an aspect was found to play an important role in determining its sentiment. This could be partly due to an overlap of tokens that are close to the aspect in both dependency and sequence graphs. However, the scatter plot shown in Figure 3, which illustrates the relationship between the two ways of measuring distance from the aspect, indicates that this is not generally the case. For example, tokens that are two steps away from the aspect in the dependency graph are, on average, five steps away in the sequence graph.

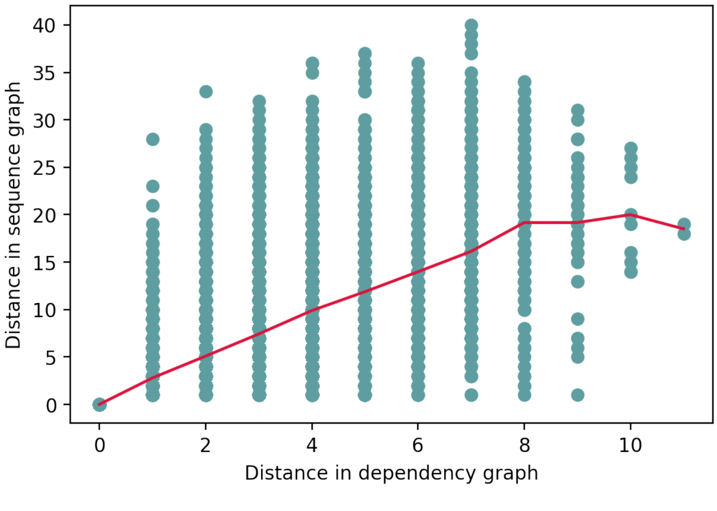

Figure 3. Relationship between the distance in dependency graph and distance in sequence graph.

We further compared the average attribution score against the distance of a token in both representations in Figure 4. We can see that the average scores for tokens that are one or two steps away from the aspect do not vary much between the two representations. Interestingly, the attribution score across the sequence graph distances is near constant for all tokens that are between one and six tokens away. On the other side, we observe a sharp decline in the attribution score for distances more than three steps away in the dependency graph. This indicates that the dependency graph distance is a better discriminator of relevant features according to their attribution score. This agrees with the previous finding that the correlation between the token attribution score and the distance of the token from the aspect was stronger for the dependency graph (-0.074) than for the sequence graph

(−0.069). We, therefore, conclude that the BERT model accounts for syntactic dependencies when performing sentiment classification.

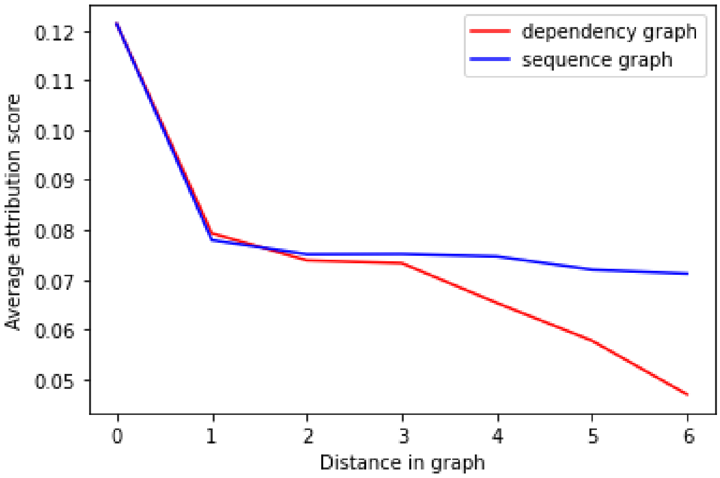

Figure 4. Average attribution score against the distance within the dependency and sequence graph.

5.4. Key Findings

The results of our analysis are in agreement with previous observations that some attention heads approximate syntactic structure by specialising to track individual dependency types [46]. Moreover, individual dependency types are often tracked by the same heads across typologically diverse languages [47]. At the same time, not all dependency types are tracked with the same robustness [48]. Prioritising certain types of dependencies over others may provide a plausible explanation as to why the fine-tuned BERT model outperformed our previous GCN-based approach [8].

Namely, two successive convolutions were performed on each word represented by its embedding following the edges in the syntactic dependency graph, hence propagating information across the graph to the second-order neighbour. This approach significantly outperformed alternative approaches, which did not take the syntactic structure into account; hence, its success was attributed to the way in which it incorporated syntactic dependencies into the logic of the neural network. However, despite their apparent value for the task of aspect-based SA, the convolution was applied to all syntactic dependencies indiscriminately—in other words, predetermined convolution across explicit syntactic dependencies. In this study, the test data suggest that the model takes into account implicit syntactic dependencies with the added flexibility of varying attention across these dependencies. The flexibility of the transformer-based approach embodied in the attention, which is used to prioritise certain types of information, including different dependency types, may hold the key to the superior performance of the transformed-based approach compared to that of the GCN-based one.

6. Conclusions

In this study, we presented an approach to fine-tuning the BERT language model for the specific task of aspect-based SA. BERT is pre-trained on a large dataset, which makes it robust with respect to the out-of-vocabulary problem and allows for fine-tuning the model for a specific NLP task by using a relatively small dataset. Our fine-tuned model achieved an accuracy of approximately 95% on a well-balanced test set. It outperformed our previous approach, which used syntactic information to guide the operation of a neural network. Our latest approach demonstrated that a BERT-based model cannot only compensate for the lack of explicit syntactic information but can, in fact, offer superior performance. Previous studies provided evidence that during the training phase BERT does learn some forms of linguistic structure [46–48]. In this study, we provide further evidence of this phenomenon in the context of aspect-based SA. Specifically, we focused on the syntactic dependencies that involve a given aspect. The evidence suggests that the model's attention is correlated with the degree of separation from an aspect, calculated as the number of steps away from the aspect in a syntactic dependency graph. This correlation was found to be stronger than the one calculated for the distance in the flat sentence representation. This brings us to the conclusion that the BERT model accounts for the syntactic dependencies when classifying the sentiment of the given aspect.

Finally, the high accuracy the model achieved in the realm of health and well-being opens up an array of possible applications in this domain [49]. When it comes to health, modern society tends to be preoccupied with inherently negative phenomena, such as diseases, injuries and disabilities [50]. However, for chronic patients, achieving a good quality of life does not necessarily imply the absence of symptoms that are associated with their medical condition. In reality, their quality of life is determined by the extent to which these symptoms are effectively managed. However, the negative sentiment associated with health symptoms a priori tends to skew the results of SA toward the negative spectrum. Previously, such an a priori bias made it difficult to measure sentiment in this domain [7]. This study provides evidence that a BERT model can be successfully fine-tuned to overcome this obstacle. The ability to accurately measure the sentiment associated specifically with signs and symptoms can support the development of systems designed to engage patients and monitor their self-management of chronic conditions remotely [51].

The aspect-based SA approach described in this study is based on an assumption that the aspects in questions are given a priori. This limitation could be addressed in future research by focusing on approaches that identify the aspects of SA automatically. Furthermore, the proposed models take individual sentences as input. Future work would focus on aggregating the sentiment related to a specific aspect across the whole document. Finally, we used pre-trained word embeddings. Further performance improvements could be gained by optimising the embeddings to reflect the underlying sentiment by providing a clear separation between positive and negative words in the vector space.

Author Contributions: Conceptualization, A.Ž., P.C. and I.S.; methodology, A.Ž., P.C. and I.S.; software, A.Ž. and P.C.; validation, A.Ž.; formal analysis, A.Ž., P.C. and I.S.; investigation, A.Ž.; resources, A.Ž.; data curation, A.Ž.; writing—original draft preparation, A.Ž.; writing—review and editing, P.C. and I.S.; visualization, A.Ž. and I.S.; supervision, P.C. and I.S.; project administration, I.S.; funding acquisition, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was part of a project funded by Cardiff University via the Vice Chancellor's International Scholarships for Research Excellence. The scholarship has been awarded to A.Ž., and her project was supervised by I.S. and P.C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and code are available at: https://github.com/ispasic/ABSA-attention_or_co-dependency (accessed on 14 March 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representation from Transformers
CNN	Convolutional neural network
GCN	Graph convolutional network
GRU	Gated recurrent unit
LSTM	Long short-term memory
NLP	Natural language processing
RNN	Recurrent neural network
SA	Sentiment analysis
SVM	Support vector machine

UMLS	Unified Medical Language System

References

1. Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167.
2. Cambria, E.; Poria, S.; Gelbukh, A.; Thelwall, M. Sentiment Analysis Is a Big Suitcase. *IEEE Intell. Syst.* **2017**, *32*, 74–80. [CrossRef]
3. Williams, L.; Arribas-Ayllon, M.; Artemiou, A.; Spasić, I. Comparing the utility of different classification schemes for emotive language analysis. *J. Classif.* **2019**, *36*, 619–648. [CrossRef]
4. Schouten, K.; Frasincar, F. Survey on Aspect-Level Sentiment Analysis. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 813–830. [CrossRef]
5. Korkontzelos, I.; Nikfarjam, A.; Shardlow, M.; Sarker, A.; Ananiadou, S.; Gonzalez, G.H. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *J. Biomed. Informat.* **2016**, *62*, 148–158. [CrossRef]
6. Žunić, A.; Corcoran, P.; Spasić, I. Improving the performance of sentiment analysis in health and wellbeing using domain knowledge. In Proceedings of the HealTAC 2020: Healthcare Text Analytics Conference, London, UK, 23–24 April 2020.
7. Žunić, A.; Corcoran, P.; Spasić, I. Sentiment analysis in health and well-being: systematic review. *JMIR Med. Informat.* **2020**, *8*, e16023. [CrossRef] [PubMed]
8. Žunić, A.; Corcoran, P.; Spasić, I. Aspect-based sentiment analysis with graph convolution over syntactic dependencies. *Artif. Intell. Med.* **2021**, *119*, 102138. [CrossRef]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Long Beach, CA, USA, 2017; Volume 30, pp. 5998–6008.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
11. Ruder, S.; Ghaffari, P.; Breslin, J.G. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 999–1005.
12. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 606–615.
13. Bao, L.; Lambert, P.; Badia, T. Attention and lexicon regularized LSTM for aspect-based sentiment analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL): Student Research Workshop, Florence, Italy, 28 July–2 August 2019; pp. 253–259.
14. Han, Y.; Liu, M.; Jing, W. Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. *IEEE Access* **2020**, *8*, 21314–21325. [CrossRef]
15. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative Study of CNN and RNN for Natural Language Processing. 2017. Available online: http://xxx.lanl.gov/abs/1702.01923 (accessed on 14 March 2022).
16. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. 2018. Available online: http://xxx.lanl.gov/abs/1803.01271 (accessed on 14 March 2022).
17. Basiri, M.E.; Nemati, S.; Abdar, M.; Cambria, E.; Acharya, U.R. ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Gener. Comput. Syst.* **2021**, *115*, 279–294. [CrossRef]
18. Zhao, M.; Yang, J.; Zhang, J.; Wang, S. Aggregated graph convolutional networks for aspect-based sentiment classification. *Inf. Sci.* **2022**, *600*, 73–93. [CrossRef]
19. Wu, H.; Zhang, Z.; Shi, S.; Wu, Q.; Song, H. Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis. *Knowl.-Based Syst.* **2022**, *236*, 107736. [CrossRef]
20. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* **2022**, *235*, 107643. [CrossRef]
21. Xiao, L.; Xue, Y.; Wang, H.; Hu, X.; Gu, D.; Zhu, Y. Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks. *Neurocomputing* **2022**, *471*, 48–59. [CrossRef]
22. Li, W.; Shao, W.; Ji, S.; Cambria, E. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* **2022**, *467*, 73–82. [CrossRef]
23. Özbey, C.; Dilekoğlu, B.; Açiksöz, S. The Impact of Ensemble Learning in Sentiment Analysis under Domain Shift. In Proceedings of the 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 6–8 October 2021; pp. 1–6.
24. Kazmaier, J.; van Vuuren, J.H. The power of ensemble learning in sentiment analysis. *Expert Syst. Appl.* **2022**, *187*, 115819. [CrossRef]
25. Luo, S.; Gu, Y.; Yao, X.; Fan, W. Research on Text Sentiment Analysis Based on Neural Network and Ensemble Learning. *Rev. D'Intelligence Artif.* **2021**, *35*, 63–70. [CrossRef]
26. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385.

27. Xu, H.; Liu, B.; Shu, L.; Yu, P. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 2324–2335.
28. Hoang, M.; Bihorac, O.A.; Rouces, J. Aspect-based sentiment analysis using bert. In Proceedings of the 22nd Nordic Conference on Computational Linguistics; Linköping University Electronic Press: Turku, Finland, 30 September–2 October 2019; pp. 187–196.
29. Li, X.; Fu, X.; Xu, G.; Yang, Y.; Wang, J.; Jin, L.; Liu, Q.; Xiang, T. Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis. *IEEE Access* **2020**, *8*, 46868–46876. [CrossRef]
30. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
31. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 72–78. [CrossRef]
32. Aygün, İ.; Kaya, B.; Kaya, M. Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic with Deep Learning. *IEEE J. Biomed. Health Informat.* **2021**, *26*, 2360–2369. [CrossRef]
33. Punith, N.; Raketla, K. Sentiment Analysis of Drug Reviews using Transfer Learning. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; IEEE: Coimbatore, India, 2021; pp. 1794–1799.
34. Koppel, M.; Schler, J. The importance of neutral examples for learning sentiment. *Comput. Intell.* **2006**, *22*, 100–109. [CrossRef]
35. Valdivia, A.; Luzón, M.V.; Cambria, E.; Herrera, F. Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Inf. Fusion* **2018**, *44*, 126–135. [CrossRef]
36. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef] [PubMed]
37. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), Online, 16–20 November 2020; pp. 38–45.
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
39. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 8024–8035.
40. Gräßer, F.; Kallumadi, S.; Malberg, H.; Zaunseder, S. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In Proceedings of the 2018 International Conference on Digital Health, Lyon, France, 23–26 April 2018; ACM: Lyon, France, 2018; pp. 121–125.
41. Drugs.com. Available online: https://www.drugs.com/ (accessed on 2 March 2022).
42. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. **2019**, arXiv:1910.01108. Available online: http://xxx.lanl.gov/abs/1910.01108 (accessed on 14 March 2022).
43. Ofek, N.; Rokach, L.; Caragea, C.; Yen, J. The Importance of Pronouns to Sentiment Analysis: Online Cancer Survivor Network Case Study. In Proceedings of the 24th International Conference on World Wide Web (WWW), Florence, Italy, 18–22 May 2015; pp. 83–84.
44. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. Captum: A Unified and Generic Model Interpretability Library for PyTorch. 2020. Available online: http://xxx.lanl.gov/abs/2009.07896 (accessed on 11 May 2022).
45. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of Machine Learning Research, Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.
46. Htut, P.M.; Phang, J.; Bordia, S.; Bowman, S.R. Do Attention Heads in BERT Track Syntactic Dependencies? **2019**, arXiv:1911.12246. Available online: http://xxx.lanl.gov/abs/1911.12246 (accessed on 14 March 2022).
47. Ravishankar, V.; Kulmizev, A.; Abdou, M.; Søgaard, A.; Nivre, J. Attention Can Reflect Syntactic Structure (If You Let It). In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL), Online, 19–23 April 2021; pp. 3031–3045.
48. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What Does BERT Look at? An Analysis of BERT's Attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 1 August 2019; pp. 276–286.
49. Spasić, I.; Özlem Uzuner.; Zhou, L. Emerging clinical applications of text analytics. *Int. J. Med Informat.* **2020**, *134*, 103974. [CrossRef] [PubMed]
50. Berg, O. Health and quality of life. *Acta Sociol.* **1975**, *18*, 3–22. [CrossRef]
51. Spasić, I.; Owen, D.; Smith, A.; Button, K. KLOSURE: Closing in on open–ended patient questionnaires with text mining. *J. Biomed. Semant.* **2019**, *10*, 1–11. [CrossRef] [PubMed]