Using Aspect-based Analysis for explainable Sentiment predictions

Thiago De Sousa Silveira[1], Hans Uszkoreit[1], and Renlong Ai[1]

Giance Technologies
{Thiago.Silveira,hans.uszkoreit,renlong.ai}@giance.ai
https://giance.ai

Abstract. Sentiment Analysis is the study of opinions produced from human written textual sources and it has become popular in recent years. The area is commonly divided into two main tasks: Document-level Sentiment Analysis and Aspect-based Sentiment Analysis. Recent advancements in Deep Learning have led to a breakthrough, reaching state-of-the-art accuracy scores for both tasks, however, little is known about their internal processing of these neural models when making predictions. Aiming for the development of more explanatory systems, we argue that Aspect-based Analysis can help deriving deep interpretation of the sentiment predicted by a Document-level Analysis, working as a proxy method. We propose a framework to verify if predictions produced by a trained Aspect-based model can be used to explain Document-level Sentiment classifications, by calculating an agreement metric between the two models. In our case study with two benchmark datasets, we achieve 90% of agreement between the models, thus showing the an Aspect-based Analysis should be favoured for the sake of explainability.

Keywords: Sentiment Analysis · Aspect-based Sentiment Analysis · Explainable Artificial Intelligence.

1 Introduction

With the advent of the Web 3.0, social media have become a rich source of subjective and opinionated data produced by real users, which is crucial for many areas of study. Sentiment Analysis is one of these areas, defined by the study of opinions, emotions, sentiments that people have towards products, services, organizations or topics [17]. Analyzing emotions is a way to understand human behaviour, making Sentiment Analysis useful in many real-world applications, such as in healthcare, finance, market and product analysis.

Research on Sentiment Analysis contains many tasks, we highlight two of them: Document-level Sentiment Analysis and Aspect-based Sentiment Analysis. Document-level Sentiment Analysis (DLSA) refers to sentiment classification models aimed at predicting a score or a polarity class for a given document [17]. Aspect-based Sentiment Analysis (ABSA) focuses on predicting sentiments towards targeted aspects in the document [13]. In this context, aspects refers to attributes, components and entities mentioned in the text that are targets of opinions written by the user. For both Document-level and Aspect-based tasks, Deep Neural Networks are the current state-of-the-art [16, 15, 12].

Deep Neural Networks are famous for their learning capabilities and are able to obtain high accuracy scores in text classification tasks, however they lack explainability. The learned weights within the architecture of a neural network is often used as a black-box in making predictions. The decision-making process for individual classifications are deficient of transparency and can be subject to bias and generalization. Consequently, calls for explainable systems have been made by the academia and government agencies [6], conceiving a new area of research called Explainable Artificial Intelligence (XAI).

There has been many attempts to make AI more explanatory and those can be grouped along several dimensions. One dimension concerns the relation between the explanatory machinery to the original AI models and algorithms. We witness approaches that add explainable functionalities into the existing AI systems; for instance, generative models attempt to add human-readable explanations to classifications [7]. Another class of XAI focuses on designing a separate system that interprets the decisions of a learned model by analyzing the relationship between contents of the input, inference activation patterns and output. As an example, LIME [10] analyzes the impact of the input perturbations in the predicted output of a model. A third class of approaches attempts to investigate alternatives to existing techniques that are more explanatory by the redefinition of the task and thus also by the form of the output.

The approach presented in this work falls in the third category and thus must not be judged as an attempt to merely improve and extend the existing technology for DLSA. We observe ABSA task as a more *explanatory alternative* to DLSA. We argue on the reasons to view the detection of pairs of aspects and opinions as explanatory and we then investigate the relation between the two in their learning and inference performance. We also discuss the advantages and disadvantages of the more demanding ABSA technology, especially the additional modelling and annotation effort.

2 Related Work

Early work on Sentiment Analysis studied sentiment of reviews using supervised learning methods via extracting features from textual data [8]. The largest limitation of these early work is that they solely focused on predicting a sentiment polarity for entire textual document (review, social media post). Aiming for a more fine-grained analysis, *Aspect-based Sentiment Analysis* was coined as a separate task than *Document-level Sentiment Analysis* [13]. In this context, ABSA focuses on predicting sentiments for each aspect (entity, attributes) within a review.

For ABSA and DLSA tasks, deep neural networks have proven to be very useful in handling sentiment classification. As of the date of publication of this work, the state of the art for DLSA and ABSA involve fine-tuning contextual word representations. Models such as XLNET [16] and BERT [4] create context-dependent bidirectional representations of words learned from real world unlabelled data; these vector representations can be easily fine-tuned into other tasks, such as document classification. The state of the art results for the DLSA task was found by fine-tuning XLNET for polarity classification [16] on movie reviews from IMDB (Accuracy 96.21%) and Yelp (Accuracy 72.2%). For ABSA task, fine-tuning BERT using double input (text and aspect) can achieve state-of-the-art results on SemEval datasets (F1 77.97% on restaurants dataset) and Sentihood datasets (Accuracy 93.6%) [12, 15].

Although deep neural networks are able to achieve high accuracy in sentiment classification tasks, these models have low levels of explainability. To a human observer, the neural network acts like a black-box and little is known about how it makes predictions [14]. The lack of transparency is a problem, because the classifications may be subject to harmful bias, generalizations or spurious correlations. For example, in the United States, a criminal risk estimation system (COMPAS) was found to make unknowingly racially biased predictions [1]. Moreover, AI systems can implicitly learn moral-sensitive bias from human texts [2]. Such problems have triggered government institutions to impose regulations on AI, such as the European Union Right to Explanation [6].

Those problems have also led the community to search for ways to create AI systems that are explainable, so-called XAI. In [5], the authors define explainability of a system as explaining data processing and representation of data within the black box model. As for explanation on data processing, LIME [10] proposes a technique to construct a local interpretable model by performing alterations in the input and checking the outputs, finding the most important features that impact the result. In addition, DeepLIFT [11] proposed a method to calculate importance scores of input features to a predicted output, a backprogation algorithm is used to compare the activation of neurons to input features. Lastly, generative methods have been proposed, such as [7], in which the authors training a model to not only predict a class, but also provide a visualization of learned weights and generate a human-readable text containing a justification.

Regarding Sentiment Analysis, few studies focus on explaining sentiment predictions. While lexicon-based and rule-based Sentiment Analysis methods are explainable by themselves, XAI methods for supervised models can generally be used in Sentiment Classification, both ABSA and DLSA tasks [18]. Although there has been few mentions about explainable sentiment analysis, such as in [3]; to our knowledge, explainability is under-researched in the area, mainly when considering the relationship between ABSA and DLSA tasks.

3 Aspect-based Sentiment Analysis as an explanation for Document-level Sentiment Analysis

Aspect-based Sentiment Analysis and Document-level Sentiment Analysis have different levels of explainability. Models for DLSA have low explainability. As DLSA only predicts a sentiment score or sentiment class for a document, it is unclear what the predicted single value represents to the end user. For instance, long documents contain multiple sentences with many, sometimes diverse, arguments and the opinions are spread around many points of discussion in the text, which we call aspects. By just predicting a unique score, DLSA neglects these multiple opinionated data and a single value prediction loses meaning and interpretability.

On a different direction, Aspect-based Sentiment Analysis has a higher level of explainability. In fact, ABSA identifies aspects in the text and their associated sentiment; thus, predicting a vector of aspect-opinion tuple for each document. In this case, an opinion can be a sentiment polarity or continuous score, or even opinion words contained from the text. Although the decision process is not transparent, ABSA models produce richer opinionated details from documents in comparison with unique valued predictions performed by DLSA models.

To exemplify, consider the user review about a restaurant: "*Although I disliked the service, the food was very delicious and the decoration is awesome!*". A document-level sentiment analysis would predict the polarity class *positive* for the review. However, an ABSA model would state that "service" is *negative*, "food" is *positive* and the "decoration" is *positive*. In this example, the classification made by the DLSA model is a general summarization and lacks the complete picture. By adding an aspect-based analysis, human understanding of the DLSA model's classification is benefited and explanation is enhanced.

It is important to state the difference regarding explainability of using ABSA in comparison with other proposed XAI methods. Many of previous proposed methods range from highlighting input features to generating a true explanation, i.e., stating in natural language why a decision was made. Actually, a sentiment analysis model that was trained on a sentiment lexicon could highlight the positive and negative sentiment words that affect the DLSA decision. However, the user would still have to read the remainder of the text to understand why, for instance, a review is classified as positive or negative. By using ABSA, on the other hand, the output contains the aspects of the object that have been associated with positive or negative sentiment. In this way, the ABSA system is explanatory by design, a human observer can understand the multiple targets of sentiment in the text, as well as their associated opinions and thus rationalize the decision made by DLSA models.

Additionally, ABSA does not only provide an output that enables the end user to understand the reasons for the overall sentiment of a document, the prediction process considers the sentiment dimensions provided by the aspects. An ABSA system does not calculate a document-level sentiment from the learned weights of positive and negative polarity indicators but makes separate decisions for every dimension of sentiment in the document. In this way, unrelated sentiment features to an aspect are disregarded, which can be helpful when the review also includes non-targeted sentiments. Further, it also helps avoiding bias that DLSA commonly suffer, such as in cases when the opinions of one aspect, maybe not even an important one, are associated to several sentiment-carrying words but another central aspect, maybe a central part or function of a product, is associated with a single sentiment word. Thus, the ABSA approach is more explanatory by design as it bring more semantic meaning into interpretation of the neural inference.

4 Framework

Given the argument that ABSA can be used for providing a deep level of detail for explicability in Sentiment Analysis, we propose a framework with the goal of identifying if an Aspect-based Sentiment Analysis model can be used as a explainable model for a DLSA model.

The proposed framework applies an evaluation methodology reviewed by [5]: the completeness of a model can be evaluated by how closely an alternative model approximates to the model under scrutiny. In our framework, ABSA can be seen as an alternative model, while the DLSA is the original model with aimed to be explained. Given a dataset of sentiment reviews, we perform predictions with DLSA models (original) and ABSA (alternative model) and we average each individual aspect-based sentiment predictions for each document of the dataset. Then, with the original predictions and the averaged alternative predictions, we calculate the agreement level between the models. If the agreement is high enough, we can say that ABSA can be used to explain DLSA.

Algorithm 1 Averaging of Aspects' sentiments

procedure AVERAGESENTIMENT(S)
    if $type ==$ "*classification*" **then**
        $countPolarity = \{\}$
        for all $s \in S$ **do**
            $countPolarity[s] += 1$
        if $countPolarity[\text{"}pos\text{"}] > countPolarity[\text{"}neg\text{"}]$ **then return** "pos"
        if $countPolarity[\text{"}neg\text{"}] > countPolarity[\text{"}pos\text{"}]$ **then return** "neg"
        if $countPolarity[\text{"}pos\text{"}] == countPolarity[\text{"}neg\text{"}]$ **then return** "neu"
        return "neu"
    else if $type ==$ "*score*" **then**
        return $\frac{1}{|S|} \sum_{i=1}^{|S|} S_i$

Formally, Let C be a collection of textual documents for which we desire to have sentiment predictions, each document $d \in C$ contains many aspects $d = \{a_1, a_2, ..., a_n\}$. Also, let ρ_{dlsa} and ρ_{absa} be two trained models for DLSA and ABSA, respectively. In this framework, $\forall d \in C$, we use the models ρ_{dlsa} and ρ_{absa} to produce sentiment predictions P_{dlsa} and P_{absa}, respectively. A given prediction $p_i \in P_{dlsa}$ represents a polarity class (positive, negative, or neutral) for $d_i \in C$, but the same prediction $p_i \in P_{absa}$ consists on a list of sentiments for each aspect in d_i ($p_i = \{s_1, s_2, ..., s_n\}$). To make the two prediction sets comparable, each $p_i \in P_{absa}$ must be averaged. Algorithm 1 is the proposed method to average the sentiment of the data, either by majority voting on the polarity of the aspects or by taking the mean of the scores.

Then, given P_{dlsa} and the averaged predictions for P_{absa}, both models can be compared by an agreement metric. Accuracy can be used as an agreement metric to calculate the percentage of documents that were classified with the same polarity by both methods. This framework shows that ABSA can be used to explain DLSA sentiment classifications if two conditions are obeyed: 1) if agreement is high enough ($> 85\%$); otherwise, the two models are not equivalent; 2) if models used for ABSA and DLSA have similar architecture and training data; elseways, the agreement could be merely by chance or by other means. To illustrate, Figure 1 shows the workflow of the framework.

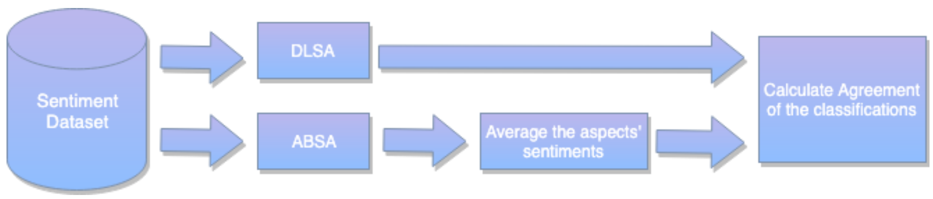

Fig. 1. A framework for agreement calculation between Aspect-based Sentiment Analysis and Document-level Sentiment Analysis

5 Case Study

To show the applicability of the framework for explainability in sentiment analysis, we make a case study applying the framework to show how ABSA predictions can be used to interpret Sentiment Analysis.

For this case study, we used two datasets from SemEval2016 Task 5, widely used for Aspect-based Sentiment Analysis [9]: 1) **Restaurants** dataset contains 2276 reviews about restaurants and 12 aspects annotated for their sentiment polarity. 2) **Laptops** dataset contains 2500 user reviews about laptops and 81 aspects that describe parts and functionality of the laptops. As for data treatment, the text was converted to lower case and grammatical errors were corrected. For the laptops dataset, we removed some of the aspects as these aspects rarely occur in the dataset. Both datasets contain three polarity classes: positive, negative and neutral.

The two datasets do not contain document-level sentiment annotations. We derived their gold standard sentiment via majority voting of the given annotated aspect sentiment for each document. Aspect-less documents or conflicted sentiments are removed. To avoid bias, a human annotator checked the document-level annotations for wrong summarizations.

The two datasets were used to train supervised models for DLSA and ABSA. We fine-tuned pre-trained BERT models [4]. We used the uncased BERT-base model with 12-layer, 12-heads, 110 million parameters for 3 epochs. For DLSA the inputs is only the document text and for the ABSA, the network receives two inputs: the document text and an aspect textual representation (either term or an aspect class), divided by a separator. To accommodate the contextual representations, we defined the input size as 128 words.

As for the experiment, we used the provided train and test from SemEval dataset. In our experiments we calculate the Accuracy and Macro-F1 for each model and the agreement between the DLSA and averaged ABSA.

5.1 Results and Discussion

Dataset	BERT DLSA		BERT ABSA		Agreement
	Accuracy	F1	Accuracy	F1	
Restaurant 2016 (Majority Vote)	85.34	72.36	87.89	73.05	90.11%
Restaurant 2016 (Reviewed)	85.68	73.57	87.89	73.05	90.63%
Laptops 2016 (Majority Vote)	81.5	65.18	81.27	65.16	89.70%
Laptops 2016 (Reviewed)	82.02	65.96	81.27	65.16	90.30%

Table 1. Agreement results for DLSA and ABSA using BERT.

The following shows the results of this case study and a further discussion. First, Table 5.1 shows the accuracy and F1 metrics for BERT models on both datasets, as well as the agreement level between BERT DLSA and the averaging of BERT ABSA predictions. The same table also contains the results for the version of the datasets reviewed by the annotation regarding the document-level sentiment polarity. The results show that BERT reaches state of the art accuracy scores. Whatsmore, the agreement between DLSA and ABSA is around 90%. Such high level of agreeement show that averaging aspect's sentiment of a document correlates with the overall sentiment of the document. In this case, we are basing the comparison of our predictions with the same pre-trained model, BERT, thus the agreement would not happen by chance.

It is also worth to analyze the results obtained by the datasets reviewed by the annotator in comparison with the version automatically made by majority voting. In this case, the datasets made by majority voting approximates to the version annotated by the user in Accuracy and F1 for the DLSA and the agreement keeps statistically the similar. Such findings have two implications: 1) DLSA and ABSA are indeed interrelated, such that using an dataset for DLSA by averaging ABSA is able to approximate the result of using datasets annotated by humans; 2) Majority voting can be used for automatically producing document-level sentiment datasets from aspects' sentiment annotations, instead of manually annotating them.

Additionally, we analyze and discuss the 10% of disagreement between the models. Table 5.1 shows the disagreement between the classes of DLSA and ABSA averaging through the framework for the Restaurants(left) and Laptops(right) dataset. Interesting, we see that the disagreement often happens regarding neutral class. Some documents classified as neutral by the DLSA method is often predicted to be positive or negative by averaging ABSA predictions. Such behaviour is reasonable as neutral classes often contains conflicted polarity information and the ABSA averaging is an approximation of the overall sentiment. Naturally, it is expected for neutral and the remaining polarity classes to be conflicted. Nevertheless, both DLSA and ABSA models have somewhat between 80%-87% of the time, thus some mistakes will certainly happen, showing that it is important to have accurate, yet consistent models.

		ABSA Averaging					
		Restaurant dataset			Laptop dataset		
		Pos	Neg	Neu	Pos	Neg	Neu
DLSA	Pos	0	10	5	0	5	10
	Neg	7	0	1	7	0	12
	Neu	20	15	0	8	16	0

Table 2. Comparison between DLSA classification and ABSA averaging for the restaurant and laptops dataset.

To clarify on the disagreement made by the involvement of neutral classes, Table 5.1 present some examples of classified documents by DLSA and ABSA. The examples chosen for this case study can show us two characteristics: 1) contradicting aspect sentiment for the disagreement sentences; and 2) detailed opinionated data makes a difference in understanding the sentiment classification. As for the former, besides cases in which the classification commits mistakes, the disagreement between DLSA and ABSA average often occurs because there are multiple contrasting sentiment within a document. However, the second characteristic is more compelling: DLSA predictions are not self explanatory. The examples given table 5.1 exemplify the argument of this work. When presented individually, predictions made by DLSA methods do not contains explanations to why it was sentiment was assigned. In fact, a human observer can only understand the DLSA sentiment predictions by analyzing the opinions associated to the aspects, thus providing explainability to the whole system.

6 Conclusion

This work has discussed that Aspect-based Sentiment Analysis is a more explanatory method than the commonly used Document-level Sentiment Analysis. ABSA provides detailed opinionated analysis of aspects and their sentiment polarity, and by design it eases interpretation by the end user, in contrast with a DLSA architectures that just predicts single sentiment class for a document, lacking details and explanations. Using ABSA is a big step in the direction of more explanatory system designs in Sentiment Analysis. To show an example of how ABSA can be used to enhance explainability, we proposed a framework to compare ABSA and DLSA models. The framework was applied in a case study in two user reviews scenarios and showing that generalizing ABSA predictions can lead to high agreement levels with DLSA models.

We can extend the discussion about the level of explanatory AI systems are required to achieve. The explanations developed in this article are shallow in

Dataset	Sentence	DLSA	ABSA	ABSA AVG
Restaurants	It is not the cheapest sushi but has been worth it every time.	pos	Food Prices=neg Food Quality=pos	neu
	Great pizza, poor service	neu	Food Quality=pos Service=neg	neu
	It is a great little place with tons of potential to be a neighbourhood joint if the service were not so impersonal and corporate-like.	pos	Restaurant=pos Service=pos	pos
Laptops	This, added with the fact that the speed of the hard drive is very slow for the money, detracts from the computer is value.	neg	Hard Disc=neg Price=neu Laptop=pos	neu
	For the price ($800!), you get a nice fast laptop, but if you ask me, it is missing some things that i feel should be automatically included.	neu	Price=neu Laptop=pos Operation Perf.=neg Design Features=neu	neu
	It is a steal when considering the specs and performance as well.	neg	Price=neg Design Features=neg Operation Perf.=neg	neg

Table 3. Classification examples for DLSA and ABSA. The sentences with red background have disagreement between DLSA and the average of ABSA's sentiments. In green, there are examples of sentences in which both methods agree.

comparison with research that aims to shed light onto the contents of a black-box model, however, no system can produce a complete explanation, due to its complexity. In the current state of XAI, we should opt for explanations that allow the user to understand the decision made by a neural network, therefore being dependent on the user. For example, for a medical diagnosis system, the explanatory output may differ depending on whether the user is a patient or a physician. Additionally, we may want to derive deeper ABSA explanations by assigning different user-adapated weights to different aspects. These two topics are worthy to be investigated in the future.

References

1. Angwin, J., Larson, J., Kirchner, L., Mattu, S.: Machine bias (Mar 2019), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Caliskan, A., Bryson, J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**, 183–186 (04 2017)
3. Clos, J., Wiratunga, N., Massie, S.: Towards explainable text classification by jointly learning lexicon and modifier terms (2017)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805

5. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (Oct 2018). https://doi.org/10.1109/dsaa.2018.00018, http://dx.doi.org/10.1109/dsaa.2018.00018
6. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**, 50–57 (2017)
7. Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8779–8788 (2018)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. pp. 79–86. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1118693.1118704, https://doi.org/10.3115/1118693.1118704
9. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 19–30 (2016)
10. Ribeiro, M.T., Singh, S., Guestrin, C.: why should i trust you?. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16 (2016). https://doi.org/10.1145/2939672.2939778, http://dx.doi.org/10.1145/2939672.2939778
11. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences (2017)
12. Sun, C., Huang, L., Qiu, X.: Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. CoRR **abs/1903.09588** (2019), http://arxiv.org/abs/1903.09588
13. Thet, T.T., Na, J.C., Khoo, C.S.: Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science **36**(6), 823–848 (2010). https://doi.org/10.1177/0165551510388123, https://doi.org/10.1177/0165551510388123
14. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable ai: A brief survey on history, research areas, approaches and challenges. International Conference on Natural Language Processing and Chinese Computing, Explainable Artificial Intelligence Workshop 2019
15. Xu, H., Liu, B., Shu, L., Yu, P.S.: BERT post-training for review reading comprehension and aspect-based sentiment analysis. CoRR **abs/1904.02232** (2019), http://arxiv.org/abs/1904.02232
16. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. CoRR **abs/1906.08237** (2019), http://arxiv.org/abs/1906.08237
17. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis : A survey. CoRR **abs/1801.07883** (2018), http://arxiv.org/abs/1801.07883
18. Zucco, C., Liang, H., Di Fatta, G., Cannataro, M.: Explainable sentiment analysis with applications in medicine. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1740–1747. IEEE (2018)